Pontificia Universidad
JAVERIANA
Cali

[VIGILADA MINEDUCACIÓN Res. 1222 de 2016]

Pontificia Universidad Javeriana Cali, Facultad de Ingeniería y Ciencias.
MAESTRÍA EN CIENCIA DE DATOS

Predicción del Porcentaje de Ahorro Energético en Edificaciones de Colombia: Un Enfoque Basado en Variables de Sostenibilidad y ASHRAE

Johan Sebastián Bolívar Sora

Jesús Rafael Gallardo Esparragoza

Director

Dra. Sandra Ramírez Buelvas

Pontificia Universidad Javeriana Cali

Codirector

Dr. Gustavo Adolfo Arteaga Botero

Pontificia Universidad Javeriana Cali

A mi familia, por ser mi refugio y mi motor en cada momento, y a mis amigos, por sus palabras de aliento y por recordarme siempre disfrutar el proceso. Gracias por estar a mi lado en este camino y hacer posible que alcance esta meta.
Johan

A mi familia, por su amor, paciencia y sacrificio que me han inspirado a dar siempre lo mejor de mí, y a mis amigos, por su apoyo constante y por ser una fuente inagotable de motivación y alegría. Este logro también es suyo.
Jesús

Índice general

Resumen

Una metodología de predicción del ahorro energético basado en la metodología ASHRAE, aplicado a tipologías de edificios en Colombia, es de gran importancia para la construcción sostenible. La implementación de estándares específicos para edificaciones en la zona ecuatorial es esencial, dadas las condiciones ambientales únicas de esta ubicación geográfica y los requerimientos de certificaciones energéticas internacionales. Actualmente, los estándares internacionales de construcción sostenible están diseñados para maximizar el aprovechamiento energético en edificaciones situadas en climas extremos, donde la dependencia de sistemas de climatización es considerable. Este enfoque amplio resulta inadecuado para regiones con condiciones climáticas favorables, como Colombia, ya que incrementa los costos de construcción y desincentiva prácticas sostenibles en tales contextos. Para abordar esta problemática, el presente estudio propone una metodología estadística que permite predecir el porcentaje y nivel de ahorro energético, utilizando variables clave para la sostenibilidad en edificaciones según la metodología ASHRAE. La aplicación de esta metodología tiene como objetivo proporcionar información que facilite el cumplimiento de métricas adaptadas a las particularidades de las tipologías de edificios y al entorno ambiental característico de Colombia.

Introducción

La construcción sostenible en países ecuatoriales enfrenta desventajas en certificaciones de sostenibilidad frente a países desarrollados. Estos entornos demandan enfoques adaptados a condiciones locales, evitando la aplicación directa de criterios de climas templados, lo que abre oportunidades para innovar en soluciones constructivas adecuadas.

Las evaluaciones actuales emplean parámetros como el consumo de energía final, el uso de energía primaria no renovable y el aislamiento térmico [1], ajustados a condiciones extremas de países con estaciones climáticas. Estos criterios no se corresponden completamente con climas de países como Colombia, Venezuela, Ecuador o Perú, donde se depende menos de sistemas de climatización artificial.

En este contexto, la eficiencia energética en edificaciones adquiere relevancia, especialmente en países en desarrollo como Colombia, donde el sector de la construcción representa una parte significativa del consumo energético. Según la *Agencia Internacional de Energía*, se prevé que América Latina y América Central contribuirán con cerca del 6 % del consumo energético mundial en edificaciones para 2050, un aumento del 40 % respecto a los niveles actuales [2]. La implementación de modelos de predicción de ahorro energético en edificios puede mejorar la eficiencia energética y reducir costos.

La asociación de ingeniería, *La American Society of Heating, Refrigerating and Air-Conditioning Engineers* (*ASHRAE*), fundada en 1959, promueve la eficiencia energética, la calidad del aire y la sostenibilidad. *ASHRAE* contribuye a la reducción del consumo energético mediante la creación de normas y estándares, que se vuelven cruciales ante el aumento de la demanda energética mundial. La metodología *ASHRAE* 90.1 2010 proporciona un marco ampliamente reconocido para el análisis y predicción del consumo de energía en edificios [3]. Esta guía se enfoca en analizar y comparar métodos para determinar cargas térmicas,

incluyendo sistemas de enfriamiento y calefacción [4].

El objetivo de esta investigación es diseñar un modelo predictivo para estimar el nivel y el porcentaje de ahorro energético en edificaciones ubicadas en Colombia. Este modelo se fundamentará en variables seleccionadas y en el consumo energético calculado de acuerdo con la metodología *ASHRAE* 90.1 2010, incorporando además las condiciones climáticas y de uso particulares del contexto colombiano.

Este estudio emplea un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de las variables y, posteriormente, modela el desempeño energético de los edificios. Concretamente, se estima (i) la categoría de ahorro energético —clasificada como *bueno*, *alto* o *muy alto*— mediante un análisis discriminante lineal robusto y (ii) el porcentaje exacto de ahorro energético a través de una regresión beta. El análisis se desarrolla en escenarios plausibles adaptados al contexto colombiano y utiliza las variables definidas en la norma ASHRAE 90.1–2010.

Los resultados podrían beneficiar a ingenieros, urbanistas y responsables de políticas en el sector de la construcción en Colombia. Pues, podría facilitar el diseño de edificaciones más eficientes, la reducción de costos energéticos y el fomento de prácticas sostenibles en construcción. También podría mejorar los procesos de certificación de edificaciones mediante parámetros ajustados a condiciones locales, impulsando la innovación en sostenibilidad con respaldo en la ciencia de datos. El modelo predictivo también es una herramienta que sirve para que los propietarios de proyectos que deseen aplicar a una certificación sostenible puedan tener mas certeza de si asumen o no con los gastos relacionados a la certificación.

Lista de figuras

Lista de tablas

1 Definición del problema

1.1. Planteamiento del problema

La reducción del consumo energético en el sector de la construcción requiere un enfoque en edificaciones de alto desempeño energético. Esto implica que los diseñadores de todas las disciplinas contribuyan desde sus respectivas áreas para desarrollar un diseño que cumpla con las expectativas de eficiencia energética establecidas por los propietarios de los proyectos.

La disparidad entre los estándares internacionales de eficiencia energética para edificaciones y las condiciones locales en países en desarrollo como Colombia representa un obstáculo significativo para la construcción sostenible. Normas como *ASHRAE*, desarrolladas para climas extremos, presentan requisitos difíciles de cumplir en contextos tropicales, donde las condiciones climáticas son moderadas y la disponibilidad tecnológica varía[5], [6]. Este desajuste plantea desafíos para la industria de la construcción local, entre ellos, la dificultad para alcanzar certificaciones internacionales, la ausencia de metodologías y herramientas adaptadas, y la escasez de profesionales especializados en prácticas sostenibles [7].

Además, la imposibilidad de cumplir con estos estándares internacionales puede incrementar los costos de los proyectos y limitar el acceso a financiamiento verde [8]. La falta de armonización entre las regulaciones locales y los estándares internacionales genera incertidumbre, lo que desincentiva la adopción de prácticas sostenibles en el sector.

La necesidad de adaptar los estándares de construcción sostenible a contextos específicos es evidente, particularmente en países con climas tropicales como Colombia. Los sistemas de certificación actuales, diseñados principalmente para climas extremos, no siempre se ajustan a las particularidades locales, situando a estos países en una desventaja competitiva en el ámbito de la sostenibilidad [9].

El Consejo Colombiano de Construcción Sostenible (CCCS) debe liderar los esfuerzos para promover métricas y criterios de certificación adecuados para países tropicales. Esto facilitará no solo una mayor competitividad internacional, sino también el desarrollo de soluciones constructivas sostenibles y adaptadas al contexto local.

Cabe destacar que las certificaciones de sostenibilidad son relativamente recientes en Colombia. *Leadership in Energy and Environmental Design* (LEED), una de las certificaciones de sostenibilidad más reconocidas a nivel mundial, se introdujo en el país hace apenas 14 años, con el primer edificio certificado en 2010 [10]. Esto subraya la necesidad de avanzar en el campo de la modelación energética.

En este contexto, es fundamental el desarrollo de metodologías de evaluación de la eficiencia energética que se adapten a las condiciones locales. Entre estas, se destaca la adaptación de modelos energéticos, la cual debe aplicarse en todas las etapas del proceso de certificación, desde la selección de materiales hasta la selección de equipos mecánicos y eléctricos [11].

1.2. Formulación del problema

¿Cómo pueden los modelos estadísticos aplicados a aspectos estructurales y operativos de los edificios, especificados por la metodología ASHRAE, ser utilizados para predecir el porcentaje y nivel de ahorro energético en edificaciones en Colombia, y cómo se correlacionan estos niveles de ahorro con las características de las edificaciones para promover una mayor eficiencia sostenible? A partir de esta pregunta de investigación, nacen las siguientes preguntas de sistematización:

1. ¿Cómo se describe la distribución y las asociaciones entre el consumo energético o porcentaje de ahorro energético y los aspectos estructurales y operativos de los edificios especificados por la metodología *ASHRAE*?

2. ¿Qué metodologías pueden ser utilizadas para predecir el porcentaje y nivel de ahorro energético en edificaciones mediante modelos estadísticos predictivos, empleando una variable de respuesta acotada entre 0 y 1, asi como métodos de agrupación y clasificación en función del nivel de ahorro energético, aplicados a los aspectos estructurales y operativos de los edificios conforme a los lineamientos especificados por la metodología *ASHRAE*?

3. ¿Cómo pueden predecirse el porcentaje y nivel de ahorro energético de un conjunto de edificios en Colombia, utilizando los aspectos estructurales y operativos de las edificaciones definidos por la metodología *ASHRAE*?

2 Objetivos del proyecto

2.1. Objetivo general

Desarrollar una metodología para predecir el porcentaje y nivel de ahorro energético en edificaciones en Colombia y su correlación con aspectos estructurales y operativos de los edificios, utilizando modelos estadísticos aplicados a los criterios especificados por la metodología *ASHRAE*, con el fin de promover una mayor eficiencia energética sostenible.

2.2. Objetivos específicos

1. Describir la distribución y las asociaciones entre el consumo energético o porcentaje de ahorro energético y los aspectos estructurales y operativos de los edificios especificados por la metodología *ASHRAE*.

2. Desarrollar una metodología para predecir el porcentaje y nivel de ahorro energético en edificaciones mediante modelos estadísticos predictivos, empleando una variable de respuesta acotada entre 0 y 1, asi como métodos de clasificación en función del nivel de ahorro energético, aplicados a los aspectos estructurales y operativos de los edificios conforme a los lineamientos especificados por la metodología *ASHRAE*.

3. Predecir el porcentaje y nivel de ahorro energético de un conjunto de edificios en Colombia, utilizando los aspectos estructurales y operativos de las edificaciones definidos por la metodología *ASHRAE*.

2.3. Resultados esperados

1. Reporte con un análisis descriptivo de la distribución y características clave del consumo energético y sus asociaciones con los aspectos estructurales y operativos de los edificios certificados en la muestra. Este análisis incluye el procesamiento de datos en relación con valores faltantes, imputación, análisis de atípicos y un análisis descriptivo univariado y multivariado entre las variables predictoras y de respuesta. Asimismo, se presentan las transformaciones de datos y categorizaciones necesarias para los análisis posteriores.

2. El informe presenta los análisis y comparaciones realizados entre diferentes métodos y modelos para determinar la metodología propuesta para predecir el porcentaje o nivel de ahorro energético en edificios en Colombia. Este documento incluye un análisis exhaustivo y comparativo de modelos que predicen el porcentaje de ahorro energético basado en los aspectos estructurales y operativos de edificios certificados. Las comparaciones consideran las métricas de desempeño de los modelos y el análisis de residuos para validar los supuestos de modelado. Adicionalmente, se incluyen análisis y comparaciones de modelos que estiman el nivel de ahorro energético, junto con un informe comparativo de las métricas de desempeño de dichos modelos.

3. Informe que presente los resultados y análisis obtenidos utilizando la metodología propuesta. Esto incluye el reporte de predicciones, las correlaciones entre el porcentaje y nivel de ahorro y las características de los edificios, aplicadas a escenarios plausibles en el contexto de edificaciones en Colombia.

3 Alcance

La norma *ASHRAE*, en particular la *ASHRAE Standard 90.1 2010* que es en el la que mas ampliamente se utiliza en Colombia, establece un marco de evaluación de la eficiencia energética en edificaciones a través del enfoque de *Building Energy Modeling* (BEM). Este método compara el desempeño energético del edificio en análisis con un modelo de referencia que cumple con los estándares mínimos de eficiencia, empleando herramientas de simulación como *EnergyPlus* y *DOE-2*. Estas herramientas se fundamentan principalmente en modelos físicos basados en ecuaciones de balance energético y principios termodinámicos, complementados ocasionalmente con parámetros empíricos para mejorar la precisión en escenarios complejos donde el modelado puramente físico resulta insuficiente. Al tratarse de modelos deterministas, los resultados de las simulaciones son consistentes cuando se emplean parámetros de entrada idénticos, generando predicciones coherentes en términos de consumo energético, temperaturas y flujos de aire.

En este contexto, el presente trabajo tiene como alcance el desarrollo de una metodología basada en modelos estadísticos para predecir el porcentaje y nivel de ahorro energético en edificaciones colombianas. La propuesta se fundamenta en las variables consideradas en la metodología *ASHRAE*, adaptándose a las condiciones climáticas y energéticas específicas de las diferentes regiones del país. Esta metodología es aplicable a diversos tipos de edificaciones (residenciales, comerciales e industriales), con el propósito de maximizar su impacto en la promoción de la eficiencia energética a nivel nacional.

Este trabajo se desarrolla bajo la limitación de no disponer de datos reales de consumo energético ni de porcentaje de ahorro de energía. En su lugar, se emplean estimaciones obtenidas a partir de simulaciones realizadas con herramientas como *EnergyPlus* y *DOE-2*, utilizadas en la metodología *ASHRAE Standard 90.1 2010*. Estas simulaciones se aplican a los 49 proyectos que conforman la

muestra, configurados con combinaciones específicas de parámetros asociados a cada proyecto. La muestra contiene un total de 44 proyectos certificados y 5 no certificados que se utilizaron para entrenar el modelo estadístico, por lo que se estimaron niveles de ahorro energético clasificados como bajo, bueno y alto. También se tomará en cuenta la métrica de la certificación LEED donde un edificio que logre un ahorro energético del 5 % o mas sobre el modelo de referencia de *ASHRAE*, cumple con lo necesario para ser un proyecto con un rendimiento energético aceptable [12].

La metodología propuesta busca aproximar el comportamiento de los modelos físicos mediante el uso de variables operativas, estructurales y climáticas, con el propósito de desarrollar una herramienta estadística que permita estimar el ahorro energético sin necesidad de realizar simulaciones físicas para cada caso particular. De este modo, la información generada puede servir como insumo para edificaciones interesadas en obtener certificaciones de eficiencia energética.

Asimismo, el alcance de la metodología está limitado a las tipologías de edificaciones y zonas climáticas representativas de Colombia, y no contempla su implementación en empresas constructoras para la optimización operativa de proyectos específicos. La propuesta se enfoca en el diseño teórico de la metodología, dejando para trabajos futuros su validación práctica y su incorporación en procesos industriales.

Como línea de trabajo futura, se propone integrar series de tiempo de consumo energético real para cada proyecto. Esto permitiría refinar los modelos estadísticos y realizar comparaciones más precisas con las estimaciones generadas por los modelos físicos utilizados en la metodología *ASHRAE Standard 90.1 2010*. A su vez, este enfoque podría sentar las bases para el desarrollo de criterios más adaptados a las características específicas de edificaciones ubicadas en zonas templadas, como es el caso de Colombia.

Esta ampliación facilitaría la validación empírica de la metodología y su aplicación en escenarios más amplios (no certificados), fortaleciendo su utilidad como herramienta para la certificación y optimización de la eficiencia energética en el sector de la construcción en Colombia.

4 Justificación

La eficiencia energética en la construcción es clave para reducir emisiones y huella de carbono, apoyando la sostenibilidad ambiental y el bienestar social. El gobierno colombiano impulsa políticas hacia la neutralidad de carbono y los objetivos de desarrollo sostenible, como parte del *Plan Nacional de Desarrollo* [13]. El *Plan Energético Nacional 2020-2050* de la *Unidad de Planeación Minero Energética* (UPME) plantea metas para el ahorro energético en edificaciones, que incluyen la implementación de tecnologías de eficiencia energética y energías alternativas [14].

En Colombia, la *Guía de Construcción Sostenible para el Ahorro de Agua y Energía en Edificaciones* establece estándares para edificios sostenibles. Este estudio integra variables climáticas y de uso específicas, permitiendo predicciones más precisas y una toma de decisiones informada [15].

La metodología *ASHRAE*, ampliamente utilizada en ingeniería de edificaciones, ofrece un marco de referencia para desarrollar un modelo que permita analizar y estimar el consumo de energía en edificios, con el objetivo de mejorar la calidad y competitividad del sector de la construcción. La implementación de modelos predictivos puede ayudar a reducir los costos energéticos en el sector [3] y contribuir significativamente a los objetivos planteados por la UPME en Colombia, promoviendo la sostenibilidad energética en el sector de la construcción [14].

Adicionalmente, obtener certificaciones en eficiencia energética, como LEED, representa un beneficio significativo para los proyectos de construcción, ya que mejora su competitividad en el mercado y refuerza su compromiso con la sostenibilidad. No obstante, los costos asociados a la certificación varían en función del área del proyecto y pueden representar una inversión considerable [16]. Contar con un modelo de predicción de ahorro energético antes de asumir estos costos proporcionaría una herramienta de prefactibilidad útil, permitiendo a los

desarrolladores evaluar la viabilidad de la certificación y detectar oportunidades de mejora en el diseño energético. Esto es especialmente relevante dado que el desempeño energético es uno de los criterios con mayor peso dentro del sistema de certificación LEED. La disponibilidad de un modelo predictivo facilitaría las etapas posteriores al pago de los costos asociados a la certificación, optimizando los recursos y aumentando la eficiencia del proceso.

Por otro lado, la Cámara Colombiana de la Construcción (CAMACOL) ha impulsado iniciativas para la recopilación de información y el desarrollo de herramientas que permitan establecer líneas base y mejores prácticas constructivas en distintas regiones del país, incluyendo el Valle del Cauca. Un modelo de predicción del ahorro energético representaría un insumo valioso en estos esfuerzos, proporcionando datos cuantificables que faciliten el diseño de estrategias orientadas a la sostenibilidad en la industria de la construcción.

Los proyectos que obtienen certificación, como LEEDE, aportan beneficios sociales, como la mejora de la calidad de vida, la promoción de la salud y el bienestar, y el fomento de la educación y la conciencia ambiental. Además, contribuyen al medio ambiente mediante la reducción de emisiones de gases de efecto invernadero, la conservación de recursos naturales y la protección del entorno local [17]. La propuesta de un modelo adaptado a las condiciones específicas de Colombia, basado en la metodología *ASHRAE*, facilitaría la obtención de certificaciones en proyectos, promoviendo la sostenibilidad en el país.

Esta investigación desarrolla un modelo de predicción energética para empresas de construcción en Colombia, basado en la metodología *ASHRAE* y adaptado a las condiciones climáticas, estructurales y operativas de las edificaciones en el país. Este modelo es útil, ya que facilita la estimación del ahorro energético potencial y promueve el desarrollo sostenible en sus dimensiones ambiental, económica y social.

Los resultados de investigación pueden tener un impacto positivo para Colombia, pues sus resultados contribuyen a las metas del UPME. También se busca fomentar ajustes en los estándares de entidades como *ASHRAE*, adaptándolos a las características de edificaciones en regiones ecuatoriales. Lo que facilitaría la obtención de certificaciones en proyectos, promoviendo la sostenibilidad en el país.

5 Marco de referencia

5.1. Marco teórico

En el ámbito de las certificaciones, existen diversos ítems que deben cumplirse. El nivel alcanzado en cada uno determina la obtención de una certificación de mayor o menor rango. Un ítem comúnmente requerido es la modelación energética, pues suele tener un peso significativo en el proceso de certificación [14].

5.1.1. Certificación LEED, metodología *ASHRAE* y modelado energético

La necesidad de reducir el consumo energético en edificaciones ha impulsado el desarrollo de normativas y herramientas que permiten evaluar, medir y certificar el desempeño energético de los proyectos constructivos. Entre las más reconocidas a nivel internacional se encuentran la certificación LEED y la normativa *ASHRAE*, que trabajan en conjunto como pilares fundamentales en la construcción sostenible.

La certificación *LEED* (Leadership in Energy and Environmental Design), desarrollada por el U.S. Green Building Council (USGBC), es un sistema de evaluación que otorga reconocimiento a las edificaciones que implementan estrategias sostenibles en aspectos como eficiencia energética, uso del agua, materiales, calidad ambiental interior y ubicación [18].

En este contexto, el estándar *ASHRAE 90.1*, emitido por la American Society of Heating, Refrigerating and Air-Conditioning Engineers, se convierte en el punto de comparación clave. Esta norma establece los requerimientos mínimos de eficiencia energética para edificaciones comerciales, determinando valores de referencia para sistemas HVAC, envolventes térmicas, iluminación y equipos

[19].

Para cumplir con los créditos de energía en LEED, los proyectos deben demostrar un ahorro energético con respecto a una línea base modelada según los criterios de *ASHRAE* 90.1.

Es en este punto donde cobra importancia el *modelo energético*. Esta herramienta consiste en una simulación computacional del comportamiento energético del edificio, considerando variables como la geometría, materiales, cargas internas, condiciones climáticas, sistemas activos y horarios de operación. A través del modelo energético, se comparan dos escenarios: el edificio propuesto con sus medidas de eficiencia y el edificio de referencia (baseline) definido por *ASHRAE*. La diferencia en consumo energético y costos operativos entre ambos define el porcentaje de ahorro alcanzado, que puede traducirse en puntos para la certificación LEED [20].

La integración de estas metodologías no solo busca validar el cumplimiento normativo, sino también fomentar el diseño y operación de edificaciones más eficientes, con menores emisiones y costos a largo plazo. En consecuencia, el modelado energético basado en *ASHRAE* 90.1 se ha consolidado como una herramienta clave para lograr certificaciones sostenibles y para la toma de decisiones informadas en la etapa de diseño.

5.1.2. Modelación energética y certificaciones

1. **La modelación energética**

   La modelación energética de edificios es una herramienta clave para desarrollar soluciones eficientes y sostenibles en el consumo energético del sector de la construcción. Este enfoque permite evaluar y predecir el rendimiento energético de los edificios, considerando la eficiencia de los sistemas de climatización, la envolvente térmica, el diseño arquitectónico y las variables climáticas [3].

   Los proyectos constructivos que emplean modelos energéticos como herramienta de toma de decisiones desde las fases de prediseño y diseño pueden implementar estrategias para resolver problemas en los edificios, tales como falta de confort térmico y visual, ventilación inadecuada, sobreconsumo de energía y elementos arquitectónicos redundantes. Estos aspectos pueden afectar negativamente la productividad, generar estrés y problemas de salud en los ocupantes [21].

   El proceso de modelación comienza con un software especializado, donde se carga información planimétrica, de ocupación, climatológica y de los sistemas del edificio (iluminación, HVAC, hidrosanitario, equipos eléctricos,

entre otros). Se ejecutan simulaciones en intervalos de tiempo específicos, y los resultados se interpretan mediante gráficos para guiar la toma de decisiones. Este proceso iterativo involucra la colaboración de todo el equipo de diseño [22].

La modelación energética considera varios aspectos fundamentales:

a) *Alcance de la modelación*: Definir el objetivo del modelo energético del proyecto, ya sea garantizar el confort, determinar el consumo del diseño actual o seleccionar los elementos de la envolvente más adecuados.

b) *Componentes de la modelación energética*: Incluir la representación detallada de la envolvente (muros, techos, ventanas, cubiertas), los sistemas de climatización (calefacción, ventilación y aire acondicionado - HVAC) y los sistemas de iluminación. También se consideran las cargas internas (ocupación, equipos) y las condiciones climáticas locales que afectan el confort en los espacios.

c) *Validación y calibración*: Realizar la validación y calibración del modelo es esencial para obtener resultados confiables, asegurando que la información modelada pueda corroborarse con datos reales.

d) *Normativas y estándares*: Aplicar normativas y estándares internacionales, como la metodología *ASHRAE* 90.1, que definen los requisitos mínimos de eficiencia energética en edificaciones.

2. **La metodología *ASHRAE* 90.1**

La metodología *ASHRAE* 90.1 es un estándar internacional que establece los requisitos mínimos de eficiencia energética para el diseño y construcción de edificios comerciales y residenciales. Proporciona pautas para la modelación energética de edificaciones con el objetivo de lograr ahorros de energía y reducir el impacto ambiental.

Dentro de la metodología *ASHRAE* se consideran varios aspectos claves:

a) *Objetivos de la metodología*: El estándar busca establecer requisitos mínimos de eficiencia energética para edificios, asegurando sistemas eficientes que proporcionen confort y seguridad a sus ocupantes.

b) *Requisitos y estándares de eficiencia energética*: La metodología especifica requisitos para aspectos críticos de eficiencia energética, como la envolvente, HVAC, iluminación, energías renovables y gestión de la energía.

c) *Análisis de ahorro energético*: La metodología utiliza una línea base de referencia para comparar el consumo anual de energía del modelo propuesto o de diseño, estableciendo los ahorros energéticos logrados.

3. **Certificaciones aplicadas a Colombia**

Para el pronóstico de ahorro energético en edificaciones en Colombia, la metodología *ASHRAE* aplica la Norma 90.1-2010 para simular el consumo energético de edificios completos. Este modelo utiliza el método de series de tiempo radiantes (RTS) de la *ASHRAE* para determinar la carga térmica en edificaciones [23]. Este método sigue los principios del balance de calor de *ASHRAE* y está alineado con programas de simulación como BLAST, DOE-2 y EnergyPlus [24].

La norma *ASHRAE*, particularmente la *ASHRAE* Standard 90.1, emplea modelos de simulación energética basados en el enfoque de Modelado de Referencia de Edificios (Building Energy Modeling, BEM). Este enfoque permite realizar evaluaciones comparativas entre el edificio en estudio y un edificio de referencia que cumple con los estándares mínimos de eficiencia energética [25].

EnergyPlus y DOE-2 son herramientas de simulación que se fundamentan principalmente en modelos físicos, utilizando ecuaciones detalladas de balance energético y principios termodinámicos para simular el comportamiento de los edificios [26]. Estos modelos físicos, en ocasiones, se complementan con parámetros empíricos que incrementan la precisión en situaciones complejas en las que un modelado físico exhaustivo resulta poco práctico.

Teóricamente, si dos edificios cuentan con idénticos parámetros de entrada en una simulación con EnergyPlus o DOE-2, los resultados deberían ser iguales o prácticamente idénticos. Esto es posible porque estos modelos son deterministas: dado un conjunto específico de condiciones y parámetros (como materiales, sistemas HVAC, ocupación, clima, entre otros), generan consistentemente la misma respuesta en términos de consumo energético, temperaturas, flujos de aire, entre otros resultados.

Este enfoque busca una mejora mínima del 5 % en eficiencia energética en nuevas construcciones respecto a una línea base. El proceso incluye la elaboración de un modelo según *ASHRAE* 90.1-2010, validado por el Green Building Certification Institute (GBCI), para evaluar el rendimiento energético del edificio e identificar áreas de optimización en consumo [27].

Esta metodología es reconocida en el ámbito de simulación energética de edificios y respaldada por estudios [28] y aplicaciones [29] que demuestran su efectividad para prever el consumo energético y detectar oportunidades de ahorro.

La metodología del estándar *ASHRAE*, aunque no es la única disponible, es ampliamente utilizada en certificaciones de sostenibilidad. Estas certificaciones promueven la construcción sostenible, y muchas constructoras optan por certificar sus edificios. Ejemplos de certificaciones en construcción sostenible incluyen *Leadership in Energy and Environmental Design* (LEED), *WELL Building Standard*, *Building Research Establishment Environmental Assessment Method* (BREEAM), EDGE [30] y, a nivel nacional, *CASA Colombia* para proyectos residenciales [31].

En Colombia, según un estudio del *Consejo Colombiano de Construcción Sostenible* (CCCS), para el año 2021 se certificaron alrededor de 700 edificaciones bajo certificaciones sostenibles, predominando LEED y EDGE. Esto contrasta notablemente con los primeros edificios certificados en 2008, que sumaban menos de 25.

Este es un mercado en notable crecimiento que beneficia a numerosos actores en la construcción. Sin embargo, la certificación LEED, la más popular en Colombia según el estado de la *Construcción Sostenible* (CCCS) [32], busca aumentar considerablemente los requisitos mínimos de ahorro para obtener puntos y lograr la certificación de edificios sostenibles [33].

Esta situación podría afectar el mercado de la construcción sostenible, ya que al dificultarse el cumplimiento de los ahorros exigidos, muchas empresas podrían optar por no buscar la certificación. Esto se relaciona con la evaluación de desempeño energético en el estándar *ASHRAE*, que no está adaptada a países cercanos a la línea del Ecuador, como Colombia.

En este contexto, la investigación para desarrollar una metodología de predicción de porcentaje y nivel de ahorro energético en edificaciones en Colombia, basado en la metodología *ASHRAE*, se clasifica como investigación aplicada con el objetivo de desarrollar y validar uno o más modelos predictivos estadísticos ajustados a edificaciones en el país.

5.1.3. Métodos y modelos estadísticos

En este estudio se propone una metodología que se aplica a una base de datos compuesta por 49 proyectos de edificios con 27 aspectos estructurales y opera-

tivos definidos por la metodología *ASHRAE*. Dado que uno de los objetivos de este trabajo es proponer un modelo para estimar el porcentaje de ahorro energético y clasificar el nivel de ahorro de energía (bajo, bueno y alto), minimizando la influencia de valores atípicos, se propone una reducción de dimensionalidad mediante un *análisis de componentes principales robusto* (PCA). Además, una *regresión beta* para predecir el ahorro energético y un método de *análisis discriminante robusto* (RDA) para clasificar los niveles de ahorro en las diferentes categorías.

1. **Análisis de Componentes Principales**

   En este estudio, la dimensionalidad de los datos se reduce de las variables cuantitativas relacionadas con iluminación, climatización y otras características de las edificaciones a cuatro componentes principales. Estas componentes principales, junto con las variables categóricas, se consideran posteriormente como variables explicativas en la implementación de los modelos posteriores para predecir porcentaje de ahorro de energía y nivel de ahorro.

   El PCA es una técnica ampliamente utilizada para reducir la dimensionalidad y detectar patrones en conjuntos de datos multivariantes. Este método transforma las variables originales en un conjunto de componentes principales ortogonales que preservan la mayor parte de la varianza de los datos [34]. Sin embargo, el PCA estándar, basado en estimadores como la media y la covarianza, es altamente sensible a la presencia de valores atípicos y a fuertes asimetrías (sesgos). Incluso una pequeña cantidad de observaciones extremas puede distorsionar significativamente los resultados, afectando la estabilidad y la confiabilidad de las conclusiones obtenidas [35].

   Dado que muchas de las distribuciones de las características de los edificios de este estudio presentan valores atípicos, esta limitación se aborda mediante un PCA robusto. Este método es una extensión del PCA tradicional que reduce la influencia de valores atípicos en la descomposición de los datos, lo cual resulta fundamental en estudios como estos donde estos valores pueden representar valores extremos legítimos y, por lo tanto, no deben excluirse arbitrariamente [36].

   a) *Análisis de Componentes Principales (PCA) Estándar* [37]

   El PCA estándar se desarrolla en los siguientes pasos: primero, las variables se centran restando su media para garantizar que tengan media cero. Luego, se calcula la matriz de covarianza (cuando las variables están en la misma escala) o la matriz de correlación (cuando las variables presentan escalas diferentes). A continuación, se obtienen

los valores propios (autovalores) y vectores propios (autovectores) de la matriz de covarianza o correlación. Los valores propios representan la varianza explicada por cada componente principal, mientras que los vectores propios determinan la dirección de los componentes principales. Finalmente, los datos se proyectan en el espacio de los componentes principales utilizando los vectores propios asociados con los mayores valores propios, siendo los primeros componentes principales los que explican la mayor proporción de la varianza.

Matemáticamente, el PCA estándar usa una matriz de p variables cuantitativas con n observaciones, centrada, $\mathbf{X} \in \mathbb{R}^{n \times p}$. El PCA busca direcciones $\mathbf{v}_1, \ldots, \mathbf{v}_k$ que maximizan la varianza proyectada:

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \operatorname{Var}(\mathbf{X}\mathbf{v}) = \arg \max_{\|\mathbf{v}\|=1} \mathbf{v}^\top \mathbf{S} \mathbf{v}$$

donde $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top \mathbf{X}$ es la matriz de covarianzas. Los vectores $\mathbf{v}_i$ son los autovectores de $\mathbf{S}$ y sus valores propios λ_i representan la varianza explicada por cada componente.

Entre las limitaciones del PCA estándar se encuentra su sensibilidad a los valores atípicos (*outliers*), dado que se basa en la media y la varianza, las cuales pueden distorsionar la orientación de los componentes principales.

b) *PCA Robusto Basado en el Método de Hubert (PcaHubert)*[38]

El PCA robusto basado en el método de Hubert (PCAHubert) es una variante del PCA estándar diseñada para reducir la sensibilidad a valores atípicos. Este método emplea técnicas robustas que minimizan el impacto de valores extremos en la orientación y cálculo de los componentes principales, proporcionando una representación más estable de la estructura subyacente de los datos.

Matemáticamente, el *PCA robusto según Hubert* (`PcaHubert`) utiliza un enfoque basado en *projection pursuit*, una técnica que examina múltiples proyecciones unidimensionales de los datos para detectar direcciones de interés (por ejemplo, donde los datos presentan mayor *outlyingness*), y selecciona aquellas que muestran estructura relevante sin estar influenciadas por outliers. La medida de outlyingness de un punto x_i respecto a una proyección $\mathbf{u}$ se define como:

$$\operatorname{Outl}(x_i) = \max_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^\top (x_i - \hat{\mu})|}{\sqrt{\mathbf{u}^\top \hat{\Sigma} \mathbf{u}}}$$

donde $\hat{\mu}$ y $\hat{\Sigma}$ son estimadores robustos de ubicación y covarianza, obtenidos mediante el método *Minimum Covariance Determinant* (MCD). El MCD busca el subconjunto de $h \leq n$ observaciones cuya matriz de covarianzas tenga el menor determinante posible, lo que garantiza máxima resistencia a outliers y un alto *punto de ruptura* (hasta 50 %). Luego de identificar este subconjunto limpio, se calcula sobre él un PCA clásico que produce las componentes robustas $\mathbf{v}_1^{\text{rob}}, \ldots, \mathbf{v}_k^{\text{rob}}$, capturando la estructura principal del conjunto sin distorsión por valores extremos.

La identificación de las variables importantes en *PCAHubert* se basa en el valor absoluto de los loadings de cada variable en los componentes principales. Las variables que emergen como las más importantes son menos susceptibles a ser sesgadas por datos anómalos, en comparación con un PCA estándar.

En este estudio, entre los escenarios analizados se encuentra, la *regresión beta* utiliza como variables explicativas las variables categóricas y las componentes del PCA robusto que explican al menos un 60 % de la varianza, denotadas como $PC = \begin{bmatrix} PC1 & PC2 & PC3 & PC4 \end{bmatrix}$. Previo a la aplicación del PCA robusto, se implementa la transformación de Box-Cox y Yeo–Johnson, y luego un escalado robusto en las variables explicativas cuantitativas para reducir la asimetría y estabilizar la varianza.

2. **Transformaciones**

En este estudio se evaluaron diversos tipos de transformaciones con el objetivo de identificar cuál resulta más adecuada según las características específicas de los datos del proyecto.

a) *Transformación de Box-Cox* [39], [40]

En este estudio, la transformación de Box-Cox se aplica a las variables cuantitativas explicativas ($X_i, i = 1, \ldots, 23$) asociadas a las características de los 49 edificios. Esta técnica de transformación de potencia estabiliza la varianza y reduce la asimetría, mejorando así la normalidad. La transformación se aplica únicamente cuando todos los valores de la variable son positivos, y el parámetro de transformación λ se selecciona de manera óptima para cada variable.

La transformación de Box-Cox para una variable X se define como:

$$X_{i,\text{Box-Cox}}(\lambda) = \begin{cases} \frac{X_i^{\lambda}-1}{\lambda}, & \text{si } \lambda \neq 0, \\ \ln(X_i), & \text{si } \lambda = 0, \end{cases}$$

donde X_i representa la variable original y λ es el parámetro de potencia ajustado para maximizar la normalidad y reducir la asimetría de X_i. Tras la aplicación de la transformación de Box-Cox, se realiza un escalado robusto para centrar y estandarizar las variables transformadas.

b) *Transformación de Yeo–Johnson*

La transformación de Yeo–Johnson extiende la de Box–Cox para cubrir todo el dominio real, incluidos los valores negativos y cero, con el fin de estabilizar la varianza y aproximar la distribución de una variable a la normalidad [41].

Sea $y \in \mathbb{R}$ y $\lambda \in \mathbb{R}$; la transformación se define como

$$T(y;\lambda) = \begin{cases} \dfrac{(y+1)^{\lambda}-1}{\lambda}, & y \geq 0,\ \lambda \neq 0, \\ \log(y+1), & y \geq 0,\ \lambda = 0, \\ -\dfrac{(-y+1)^{2-\lambda}-1}{2-\lambda}, & y < 0,\ \lambda \neq 2, \\ -\log(-y+1), & y < 0,\ \lambda = 2. \end{cases}$$

La función resultante está bien definida para cualquier valor real de y y, al igual que Box–Cox, el parámetro λ suele estimarse por máxima verosimilitud para maximizar la aproximación a la normalidad.

3. **Escalamiento**

Con el fin de hallar la forma de estandarización que mejor se adapte a la naturaleza de los datos y, al mismo tiempo, facilite proyecciones estables en el PCA, se ensayaron tres métodos de escalamiento. Para cada uno se evaluó su impacto sobre (i) la simetría de las variables transformadas y (ii) la varianza explicada por los primeros componentes.

a) *Escalado robusto (Mediana–MAD)*

Tras aplicar las transformaciones previas, cada variable x se centró en su mediana y se dividió por su Desviación Absoluta Mediana (MAD):

$$x_i^{\text{esc}} \;=\; \frac{x_i - \text{Med}(x)}{\text{MAD}(x)}, \qquad \text{MAD}(x) \;=\; \text{Med}\big(\,|x_i - \text{Med}(x)|\,\big).$$

Este procedimiento disminuye drásticamente la influencia de valores

extremos y tiende a producir distribuciones más simétricas, lo que favorece la interpretación de los ejes principales.

b) *Escalado por rango intercuartílico (IQR)*
El segundo enfoque emplea la amplitud intercuartílica:

$$x_i^{\text{esc}} = \frac{x_i - \text{Med}(x)}{IQR(x)}, \qquad IQR(x) = Q_3 - Q_1.$$

Al basarse en los cuartiles (Q_1 y Q_3), este método es también robusto a valores atípicos, aunque suele ser algo menos sensible que la MAD en presencias de colas muy pesadas [42].

c) *Escalado estándar (z-score)*
Como referencia se incluyó la normalización clásica:

$$x_i^{\text{esc}} = \frac{x_i - \mu(x)}{\sigma(x)},$$

donde $\mu(x)$ y $\sigma(x)$ son la media y la desviación típica muestrales. Aun cuando es sensible a outliers, este z-score sirve de línea base para cuantificar la mejora aportada por los métodos robustos anteriores.

4. **Regresión Beta** [43]

La regresión beta es adecuada para modelar variables continuas acotadas entre 0 y 1, como el porcentaje de ahorro de energía analizado en este estudio. En este caso, la media de la distribución beta se vincula a un predictor lineal mediante una función de enlace logit, probit o complemento log-log, y dicho predictor se compone de dos bloques de covariables. En el primer escenario, se utilizan componentes principales robustas que condensan las variables cuantitativas relacionadas con el diseño y la operación de los edificios. En el segundo, se incluyen estas componentes principales junto con variables categóricas recodificadas como indicadores binarios (por ejemplo, tipo de uso, zona climática o presencia de sistemas de gestión), las cuales se incorporan sin categoría de referencia pero con la restricción de que sus coeficientes sumen cero, garantizando así la identificabilidad del modelo. En el tercer escenario, se emplean directamente las variables cuantitativas transformadas (sin reducción mediante PCA) junto con variables dummy derivadas de las variables categóricas Tipología (uso principal del proyecto) y Ciudad / Zona climática.

a) *Modelo*

Sea Y_i el porcentaje de ahorro energético del edificio i, una variable continua acotada en el intervalo $(0, 1)$. Se modela Y_i como una variable aleatoria con distribución Beta, donde la media $\mu_i = \mathbb{E}[Y_i]$ se relaciona con un predictor lineal η_i mediante una función de enlace $g(\cdot)$:

$$Y_i \sim \text{Beta}(\mu_i, \phi), \qquad g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta},$$

Aquí, ϕ representa el parámetro de precisión de la distribución; $\mathbf{x}_i$ es el vector de predictores que incluye tanto componentes principales robustas como variables categóricas recodificadas en indicadores binarios sin categoría de referencia, bajo la restricción de que sus coeficientes sumen cero para garantizar la identificabilidad; y $g(\cdot)$ es la función de enlace. Para evaluar la sensibilidad del modelo al tipo de enlace, se ensayaron tres funciones comúnmente utilizadas:

$$\begin{aligned}
\text{Logit:} \quad & g(\mu) = \log\frac{\mu}{1-\mu}, && g^{-1}(\eta) = \frac{1}{1+e^{-\eta}}, \\
\text{Probit:} \quad & g(\mu) = \Phi^{-1}(\mu), && g^{-1}(\eta) = \Phi(\eta), \\
\text{Complemento–log–log:} \quad & g(\mu) = \log\big[-\log(1-\mu)\big], && g^{-1}(\eta) = 1 - e^{-e^{\eta}},
\end{aligned}$$

donde $\Phi(\cdot)$ denota la función de distribución acumulada de una normal estándar. El enlace logit se adopta como referencia debido a su interpretación directa en términos de *odds*, mientras que las funciones *probit* y *cloglog* permiten explorar el comportamiento del modelo ante posibles asimetrías y diferencias en los extremos de la distribución.

b) *Supuestos*

- Para garantizar la validez del modelo de regresión beta, es necesario verificar ciertos supuestos. En primer lugar, la variable dependiente $\mathbf{Y}$ debe estar estrictamente dentro del intervalo $(0, 1)$, tal como lo exige la distribución beta. Por tanto, se debe comprobar que todos los valores de $\mathbf{Y} = (Y_j)_{j=1}^n$ se encuentren en ese rango. En caso de que existan valores extremos iguales a 0 o 1, se aplica la transformación recomendada por Smithson y Verkuilen [44], definida como:

$$Y_{j,\text{new}} = \frac{Y_j \times (n-1) + 0{,}5}{n},$$

donde $j = 1, \ldots, n$ y $n = 49$ es el número de observaciones o pro-

yectos en este estudio. Esta transformación desplaza ligeramente los valores extremos hacia el interior del intervalo, permitiendo así el uso adecuado del modelo beta. Esta técnica ha sido ampliamente validada para el análisis de proporciones y porcentajes.

- Se asume independencia entre las observaciones, condición que se considera satisfecha dado el diseño del estudio y la naturaleza no longitudinal de los datos.
- El modelo presupone una relación lineal entre la media del porcentaje de ahorro energético y las variables explicativas, en la escala de la función de enlace. Este supuesto puede evaluarse mediante gráficos de residuos (de deviance o de Pearson). En caso de detectar no linealidades, se podrían introducir transformaciones o términos no lineales en el modelo. Para explorar distintas estructuras funcionales, se comparan en este estudio tres funciones de enlace: logit, probit y complemento log–log (cloglog), con el fin de identificar la opción que mejor se ajuste a los datos y cumpla los supuestos del modelo.
- El supuesto de homocedasticidad (varianza constante de los errores en la escala del predictor lineal) se verifica mediante la prueba de Breusch–Pagan. La hipótesis nula establece homocedasticidad, mientras que la alternativa indica la presencia de heterocedasticidad. Se utiliza un nivel de significancia de 0,05 para la decisión estadística.
- También se evalúa si los errores en la escala de la función de enlace siguen una distribución aproximadamente normal. Para ello, se emplea la prueba de Shapiro–Wilk, cuya hipótesis nula plantea que los residuos son normales. Un valor-p menor a 0,05 sugiere desviación de la normalidad. Aunque este supuesto no es esencial para la estimación por máxima verosimilitud en modelos beta, sí puede afectar la validez de inferencias posteriores, como pruebas de hipótesis o construcción de intervalos.
- Otro supuesto importante es la ausencia de multicolinealidad excesiva entre las variables explicativas. La colinealidad puede afectar la estabilidad y precisión de las estimaciones. Este aspecto se examina mediante la matriz de correlaciones de Kendall entre las variables cuantitativas. Además, en los modelos que incorporan variables dummy sin categoría de referencia, se impone la restricción de que los coeficientes sumen cero, lo cual evita la colinealidad. Cuando se utilizan componentes principales robustas

(PCA robusto) como predictores, este supuesto también se cumple, dado que las componentes son ortogonales entre sí, es decir, no están correlacionadas por construcción.

- Aunque el uso de PCA reduce la colinealidad, esto no elimina la posibilidad de que ciertas observaciones ejerzan una influencia desproporcionada en el modelo. Por ello, es fundamental examinar la presencia de valores atípicos o puntos influyentes. Esta verificación se realiza a través de gráficos de residuos estandarizados, que permiten identificar observaciones con alta influencia potencial en los resultados del modelo.

c) *Métricas para Evaluar el Rendimiento del Modelo*

Algunas de las métricas utilizadas para evaluar y comparar el desempeño de modelos de regresión beta incluyen el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE), el Coeficiente de Determinación Pseudo-R^2, la log-verosimilitud, así como los criterios de información AIC (Criterio de Información de Akaike) y BIC (Criterio de Información Bayesiano).

- En este estudio, para comparar el desempeño predictivo de distintos modelos de regresión beta bajo validación cruzada, se utiliza el Error Absoluto Medio (MAE). Esta métrica es menos sensible a valores atípicos que el MSE y proporciona una estimación robusta del rendimiento medio del modelo, especialmente en el contexto de proporciones o tasas. El MAE se define como:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \,,$$

donde y_i representa el valor observado del porcentaje de ahorro energético, $\hat{y}_i$ es el valor estimado por el modelo, y n es el número total de edificios. Valores menores de MAE indican mejor desempeño predictivo.

- El pseudo-R^2 permite evaluar la proporción de variabilidad explicada por el modelo en comparación con un modelo nulo (sin predictores, sólo con intercepto). Una formulación común es:

$$\text{Pseudo-}R^2 = 1 - \frac{\log L_{\text{modelo}}}{\log L_{\text{nulo}}},$$

donde $\log L_{\text{modelo}}$ es la log-verosimilitud del modelo ajustado, y $\log L_{\text{nulo}}$ es la log-verosimilitud del modelo nulo. Aunque no es

directamente comparable con el R^2 de la regresión lineal, valores más cercanos a 1 indican mejor ajuste.

- La log-verosimilitud ($\log L$) mide qué tan probable es observar los datos dados los parámetros estimados por el modelo. Cuanto mayor sea este valor (más cercano a cero en modelos negativos), mejor es el ajuste del modelo. Esta métrica sirve de base para calcular otros indicadores como el AIC y BIC.
- El AIC (Akaike Information Criterion) y el BIC (Bayesian Information Criterion) penalizan la complejidad del modelo para evitar el sobreajuste. Se definen como:

$$\text{AIC} = -2\log L + 2k, \qquad \text{BIC} = -2\log L + k\log(n),$$

donde k es el número de parámetros estimados y n el número de edificios. Ambos criterios penalizan modelos con mayor número de parámetros, pero el BIC lo hace más severamente. Al comparar modelos, menores valores de AIC y BIC indican mejor desempeño relativo.

- En este trabajo, la comparación entre modelos (por ejemplo, con diferentes funciones de enlace o conjuntos de predictores) se realiza utilizando de forma conjunta el MAE para evaluar el error predictivo, el pseudo-R^2 para estimar capacidad explicativa, y los criterios AIC, BIC y log-verosimilitud para evaluar el compromiso entre ajuste y complejidad del modelo. En todos los casos, se considera que un modelo es superior cuando presenta MAE más bajo, pseudo-R^2 más alto, y menores valores de AIC y BIC. La elección final se basa en un equilibrio entre capacidad predictiva, parsimonia y ajuste global del modelo.

d) *Variables más Importantes*

Para identificar las variables más importantes que contribuyen a la estimación del porcentaje de ahorro de energía en el modelo de regresión beta, se selecciona el modelo que incluye solo variables significativas, cumple los supuestos distribucionales y minimiza el MAE. En este modelo, la importancia de cada variable se determina a partir de dos criterios principales: El valor absoluto de sus coeficientes, ya que los coeficientes de mayor magnitud representan una influencia más fuerte en la predicción del ahorro de energía; y su coherencia práctica, es decir, el grado en que el efecto de la variable es razonable y consistente con el contexto operativo. Este enfoque permite identifi-

car no solo las variables estadísticamente significativas, sino también aquellas que tienen una relevancia práctica directa en la mejora del ahorro energético.

5. **Análisis Discriminante Robusto (RDA)**

   El *análisis discriminante robusto* (RDA) es una extensión del a*análisis discriminante lineal* (LDA), diseñada para ofrecer mayor resistencia frente a valores atípicos y distribuciones sesgadas. A diferencia del LDA tradicional, que se basa en estimadores clásicos de media y covarianza sensibles a observaciones extremas, el RDA utiliza estimadores robustos que reducen el efecto de estos valores sobre la función discriminante, mejorando así la estabilidad y precisión del modelo de clasificación.

   En este estudio utilizamos *Regularized Discriminant Analysis* (RDA) para evaluar hasta qué punto las características de diseño y operación de los edificios permiten predecir variables de desempeño energético bajo dos escenarios de respuesta:

   - *Escenario binario ($k = 2$).* La respuesta indica si el edificio obtiene o no una *certificación energética* (ej. *ASHRAE*): proyectos certificados ($y = 1$) vs. no certificados ($y = 0$).
   - *Escenario ordinal ($k = 3$).* La respuesta es el *nivel de ahorro energético*, categorizado en tres niveles (*bajo*, *medio*, *alto*) a partir del porcentaje de ahorro observado.

   A continuación, se detallan los componentes y el modelo del RDA [45], [46].

   a) *Modelos LDA y RDA*

   El Análisis Discriminante Lineal (LDA) asume que cada clase k proviene de una distribución normal multivariada con media $\boldsymbol{\mu}_k$ y *matriz de covarianza común* $\boldsymbol{\Sigma}$. La regla de clasificación de Bayes (ignorando constantes comunes) es

   $$\delta_k^{\text{LDA}}(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \tfrac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln \pi_k,$$

   donde π_k es la proporción (a priori) de la clase k.

   El Análisis Discriminante Cuadrático (QDA) relaja el supuesto de covarianza común y permite una matriz $\boldsymbol{\Sigma}_k$ distinta en cada clase; la regla resultante tiene términos cuadráticos en $\mathbf{x}$.

b) *Interpolación y regularización en RDA*

RDA introduce *dos hiperparámetros* $\gamma, \lambda \in [0,1]$ para interpolar entre QDA y LDA y, simultáneamente, aplicar *shrinkage* hacia matrices diagonales que reducen la varianza de los estimadores de covarianza cuando la muestra es pequeña o los predictores están altamente correlacionados. Partiendo de las estimaciones clásicas de covarianza por clase $\widehat{\boldsymbol{\Sigma}}_k$ y de la covarianza agrupada $\widehat{\boldsymbol{\Sigma}}_{\text{pool}}$, RDA construye una versión regularizada

$$\widehat{\boldsymbol{\Sigma}}_k^{(\gamma,\lambda)} = (1-\lambda)\Big[(1-\gamma)\,\widehat{\boldsymbol{\Sigma}}_k + \gamma\,\widehat{\boldsymbol{\Sigma}}_{\text{pool}}\Big] + \lambda\,\text{diag}\Big((1-\gamma)\,\widehat{\boldsymbol{\Sigma}}_k + \gamma\,\widehat{\boldsymbol{\Sigma}}_{\text{pool}}\Big),$$

donde $\text{diag}(\cdot)$ retiene sólo los elementos diagonales (variancias). Así:

- $\gamma = 0$ y $\lambda = 0$ recupera QDA puro $\big(\widehat{\boldsymbol{\Sigma}}_k^{(0,0)} = \widehat{\boldsymbol{\Sigma}}_k\big)$.
- $\gamma = 1$, $\lambda = 0$ aproxima LDA (covarianza común).
- $\lambda \to 1$ reduce las covarianzas cruzadas (modelo más cercano a Naive Bayes gaussiano).

Sustituyendo $\widehat{\boldsymbol{\Sigma}}_k^{(\gamma,\lambda)}$ en la regla discriminante gaussiana se obtiene la función de clasificación de RDA.

En este trabajo, la rejilla evaluada varía γ, λ no se exploran los casos "puros" $\gamma = 0$ (QDA), $\gamma = 1$ (LDA) ni $\lambda = 0, 1$. Esto fuerza a todos los modelos a poseer cierta mezcla y cierto grado de *shrinkage*, mejorando la estabilidad de las matrices de covarianza con la muestra disponible.

c) *Métricas de Desempeño*

1) *Escenario ($k = 2$), clasificación «Certificación/ No certificación»*:
En adelante se trabaja con dos clases: sí ($y = 1$, proyecto *certificable*) y no ($y = 0$). Sean $\text{VP}, \text{FP}, \text{FN}, \text{VN}$ los conteos habituales de la matriz de confusión:

$$\begin{pmatrix} \text{VP} & \text{FN} \\ \text{FP} & \text{VN} \end{pmatrix}$$

Donde:

- VP: verdaderos positivos ($y = 1$ predicho correctamente).
- FP: falsos positivos ($y = 0$ clasificado erróneamente como 1).
- FN: falsos negativos ($y = 1$ clasificado erróneamente como 0).
- VN: verdaderos negativos ($y = 0$ predicho correctamente).

Con estas cantidades se calculan las métricas empleadas en este trabajo:

- Sensibilidad (o *Recall*)

$$\text{Sens} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Proporción de proyectos certificables correctamente detectados.

- Especificidad
$$\text{Spec} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$
Capacidad de evitar auditorías innecesarias.
- Balanced Accuracy
$$\text{BA} = \tfrac{1}{2}\big(\text{Sens} + \text{Spec}\big)$$
Promedia los aciertos en ambas clases para mitigar desbalance.
- Accuracy (exactitud global)
$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$
- Precisión (*Positive Predictive Value*)
$$\text{Prec} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$
Fracción de las predicciones positivas que realmente pertenecen a la clase certificada; componente crítico de la curva PR.
- AUC-ROC
Área bajo la curva *Receiver Operating Characteristic*; integra la trayectoria $\big(\text{Sensibilidad},\ 1-\text{Especificidad}\big)$ al variar el umbral de decisión.
- AUC-PR
Área bajo la curva *Precision–Recall*, donde cada punto corresponde a Recall, Precisión. Resulta más informativa que la ROC cuando la clase positiva es minoritaria porque penaliza explícitamente los falsos positivos mediante la Precisión y enfatiza la recuperación de casos certificados.

Los costos de error que se establecen $C_{\text{FN}} : C_{\text{FP}} \simeq 2{:}1$; por ello se utiliza la métrica siguiente para $\sqrt{2}$ (en la metodología se retoma este tema):

$$F_{\sqrt{2}} = \frac{(1+\sqrt{2})\,\text{Prec}\,\text{Sens}}{\sqrt{2}\,\text{Prec} + \text{Sens}},$$

donde el factor $\beta = \sqrt{2}$ hace que la Sensibilidad pese el doble

que la $\mathrm{Precisin}$, alineándose con la penalización económica (FN cuesta el doble que FP).
El procedimiento de diagnóstico final del modelo RDA se realiza construyendo:

a′ *Matriz de confusión* 2×2. Se obtienen los conteos VP, FP, FN, VN después de comparar valores reales y predichos.

b′ *Curva ROC binaria.* Se representa la trayectoria $\big(1-\mathrm{Spec}(t), \mathrm{Sens}(t)\big)$ al variar $t \in [0,1]$. El área bajo la curva se denota $\mathrm{AUC}_{\mathrm{ROC}}$ y resume la capacidad de ranking *independiente del umbral*.

c′ *Curva Precision–Recall.* Más informativa que la ROC cuando la clase positiva es minoritaria. Valores altos indican que el modelo mantiene alta precisión incluso con umbrales bajos.

2) *Escenario ($k = 3$), clasificación «Bajo/Medio/Alto»*:
Para el problema de tres clases $\mathcal{Y} = \{\text{Bajo (B)}, \text{Medio (M)}, \text{Alto (A)}\}$ las métricas se calculan con la estrategia *one–vs–rest*: se aísla cada categoría $c \in \mathcal{Y}$ y se compara contra la unión de las dos restantes (rest $= \mathcal{Y} \setminus \{c\}$). De este modo se obtienen Sens_c, Spec_c y Prec_c individuales. Los valores *macro* se construyen promediando los tres resultados a partes iguales:

$$\mathrm{Sens}_{\mathrm{macro}} = \frac{1}{3}\Big(\mathrm{Sens}_{\mathrm{B}} + \mathrm{Sens}_{\mathrm{M}} + \mathrm{Sens}_{\mathrm{A}}\Big)$$

$$\mathrm{Spec}_{\mathrm{macro}} = \frac{1}{3}\Big(\mathrm{Spec}_{\mathrm{B}} + \mathrm{Spec}_{\mathrm{M}} + \mathrm{Spec}_{\mathrm{A}}\Big)$$

$$\mathrm{BA}_{\mathrm{macro}} = \frac{1}{3}\Big(\tfrac{1}{2}(\mathrm{Sens}_{\mathrm{B}}+\mathrm{Spec}_{\mathrm{B}})+\tfrac{1}{2}(\mathrm{Sens}_{\mathrm{M}}+\mathrm{Spec}_{\mathrm{M}})+\tfrac{1}{2}(\mathrm{Sens}_{\mathrm{A}}+\mathrm{Spec}_{\mathrm{A}})\Big)$$

$$F_{\sqrt{2}}^{\mathrm{macro}} = \frac{1}{3}\sum_{c\in\{\mathrm{B,M,A}\}} \frac{(1+\sqrt{2})\,\mathrm{Prec}_c\,\mathrm{Sens}_c}{\sqrt{2}\,\mathrm{Prec}_c + \mathrm{Sens}_c}$$

donde $\sqrt{2}$ es el parámetro β que da a la *Sensibilidad* el doble de peso que a la *Precisión*, coherente con la relación de costos que posteriormente se discutirá en la metodología $C_{\mathrm{FN}} : C_{\mathrm{FP}} \simeq 2{:}1$.
Para realizar el diagnóstico final del modelo RDA, se construyen los siguientes elementos:

a′ *Matriz de confusión* 3×3. Resume los aciertos y errores para Bajo, Medio y Alto aplicando el mismo corte t^* a la proba-

bilidad de la clase ganadora. Permite identificar confusiones frecuentes (p. ej. A $\rightarrow$ M).

b′ *Curvas ROC* one-vs-rest. Para cada $c \in \{\mathrm{B}, \mathrm{M}, \mathrm{A}\}$ se traza la curva ROC comparando la probabilidad $\hat{p}_c$ frente a la etiqueta binaria $y_c = \mathbb{1}\{Y = c\}$. El área bajo la curva se promedia:

$$\mathrm{AUC}_{\mathrm{ROC}}^{\mathrm{macro}} = \tfrac{1}{3}\big(\mathrm{AUC}_{\mathrm{B}} + \mathrm{AUC}_{\mathrm{M}} + \mathrm{AUC}_{\mathrm{A}}\big).$$

c′ *Curvas Precision–Recall* one-vs-rest. Se obtienen análogamente y se reporta $\mathrm{AUC}_{\mathrm{PR}}^{\mathrm{macro}}$, más informativa cuando el número de observaciones por clase es reducido.

Estas visualizaciones permiten revisar si el modelo distingue correctamente los niveles de ahorro y verifican que la optimización con $F_{\sqrt{2}}^{\mathrm{macro}}$ y la matriz de costos $C_{\mathrm{FN}} : C_{\mathrm{FP}} = 2{:}1$ conduce a un equilibrio adecuado entre *recuperar* casos relevantes (Medio/Alto) y *evitar* sobre-estimaciones.

d) *Variables más Importantes*

A continuación se explica el criterio para revisar la importancia de las variables considerando el caso de certificación y no certificación.

El comando `varImp()` aplicado al objeto `final_rda` extrae del modelo de *Regularized Discriminant Analysis* las medias de cada predictor x_j dentro de cada clase k (μ_{kj}). A partir de ellas calcula la media global $\bar{\mu}_j$ y, para cada predictor, la suma de las desviaciones absolutas $\sum_k |\mu_{kj} - \bar{\mu}_j|$.

Con dos clases (certificado y no certificado) esta suma coincide con la diferencia absoluta $|\mu_{1j} - \mu_{2j}|$; por ello el *score* "Overall" es, esencialmente, la distancia entre las medias de "Si" y "No" tras las transformaciones de Yeo–Johnson, centrado y escalado. El procedimiento es univariado: mide cuánto se desplaza cada variable a lo largo del eje que separa las clases, sin considerar covarianzas ni la regularización γ–λ.

El gráfico que se genera para estudiar la importancia de las variables es un diagrama de barras horizontal donde la altura de cada barra representa el valor `Overall` (re-escalado si se usa `scale = TRUE`). Las barras más largas indican predictores cuyas medias difieren marcadamente entre las clases y, por tanto, poseen mayor capacidad discriminante individual. Barras cortas señalan variables con medias casi idénticas, cuya aportación univariada al criterio lineal/cuadrático es

débil; no obstante, podrían adquirir relevancia combinada a través de la matriz de covarianza que RDA modela.

En la interpretación se tiene presente que el ranking no capta interacciones ni efectos no lineales: una importancia baja no implica irrelevancia absoluta, sino menor separación de medias cuando se examina el predictor de forma aislada.

6. **Validación cruzada y LOOCV** [47]

   La validación cruzada K-Fold es una técnica ampliamente utilizada en estadística y aprendizaje automático para evaluar el rendimiento de un modelo de manera robusta y generalizable. Esta metodología consta de dos etapas principales: *división del conjunto de datos* e *iteraciones de entrenamiento y prueba*.

   En la primera etapa, el conjunto de datos se divide en K subconjuntos (o *folds*) de tamaño similar. En la segunda etapa, se realizan K iteraciones, donde en cada una de ellas se reserva uno de los folds como conjunto de prueba y se utilizan los $K-1$ restantes como conjunto de entrenamiento. El modelo se ajusta con los datos de entrenamiento y se evalúa con el fold reservado. Este procedimiento se repite hasta que cada fold ha sido usado exactamente una vez como conjunto de prueba. Al finalizar, se promedian las métricas de rendimiento obtenidas en cada iteración, proporcionando así una estimación confiable del desempeño del modelo sobre datos no observados.

   Esta técnica es especialmente adecuada cuando se dispone de conjuntos de datos limitados, como en este estudio, que cuenta con solo 49 proyectos de edificios. Su aplicación permite maximizar el uso de los datos disponibles, mejorar la estabilidad de las estimaciones y reducir el riesgo de sobreajuste al modelo.

   En este trabajo se implementa una validación cruzada tanto para la regresión beta como para el RDA.

   El *Leave–One–Out Cross-Validation* (LOOCV) puede verse como el caso límite de la validación K-Fold cuando $K = n$; es decir, cada observación se emplea una única vez como conjunto de prueba mientras las $n-1$ restantes constituyen el entrenamiento.

   En este estudio $n = 49$ (proyectos), lo que implica entrenar y evaluar el modelo 49 veces y obtener un vector de 49 predicciones *fuera de muestra* que

- *Maximizan el uso de los datos disponibles.* Cada ajuste aprovecha el 98 % de la información $(48/49)$, aspecto crucial cuando el tamaño muestral es reducido.
- *Sesgo mínimo.* Es mínimo porque, en cada iteración, la observación evaluada no participa en el entrenamiento.
- *Permiten construir curvas ROC/PR y métricas agregadas* sobre las 49 probabilidades *hold-out*, ofreciendo un diagnóstico más estable que el derivado de un único conjunto de prueba fijo.
- *Evitan fuga de información (*information leakage*).* Las fases de preprocesado y la aplicación del umbral t^* se realizan dentro de cada ciclo, manteniendo aislada la muestra excluida.
- *Tienen un costo computacional mayor* que un K-Fold con $K \ll n$; sin embargo, con 49 proyectos y un modelo RDA la carga (49 ajustes) es perfectamente asumible.

5.2. Antecedentes

El estudio de técnicas de predicción de consumo energético en edificaciones ha ganado relevancia debido a la necesidad de optimizar el uso de la energía en diversos contextos arquitectónicos. La literatura actual abarca una variedad de enfoques metodológicos que incluyen desde modelos estadísticos y técnicas de minería de datos hasta simulaciones detalladas de comportamiento térmico. Esta sección revisa estudios representativos que exploran dichas técnicas y su aplicabilidad, proporcionando un marco teórico que fundamenta la construcción de modelos predictivos efectivos y adaptables a distintos tipos de edificaciones y contextos climáticos.

En el contexto del diseño y evaluación del desempeño energético de edificaciones sostenibles, las certificaciones como *LEED (Leadership in Energy and Environmental Design)* han cobrado gran relevancia a nivel internacional y en Colombia. LEED promueve prácticas sostenibles en todas las etapas del ciclo de vida de los edificios, exigiendo la verificación del ahorro energético como uno de los pilares de su sistema de puntuación. Para validar estos ahorros, se emplean metodologías técnicas rigurosas, dentro de las cuales la normativa *ASHRAE 90.1*, especialmente en sus versiones 2007 y 2010, constituye el estándar de referencia. Esta norma establece los requerimientos mínimos de eficiencia energética en edificaciones nuevas o existentes. La interrelación entre ambas herramientas

permite que los proyectos certificados bajo LEED realicen simulaciones energéticas comparativas entre un caso base (baseline) definido por *ASHRAE* y una propuesta eficiente (proposed), lo que proporciona una base técnica y objetiva para estimar el *Porcentaje de Ahorro Energético (PAH)*. Este vínculo normativo entre LEED y *ASHRAE* es esencial para garantizar que las decisiones de diseño en proyectos sostenibles estén respaldadas por modelos energéticos robustos y consistentes [48].

En [3], se presentan diversas técnicas de predicción de consumo energético en edificaciones, explicando sus principios teóricos y las etapas involucradas en la construcción de un modelo de predicción, así como las métricas de evaluación. Además, examina las fortalezas y debilidades de cada técnica y ofrece criterios para seleccionar técnicas y métricas de evaluación adecuadas según las características del caso de estudio.

La investigación sigue un diseño bibliométrico para identificar y analizar los artículos más relevantes sobre demanda energética en edificios. Este análisis permite detectar tendencias en la aplicación de técnicas de predicción de consumo energético en edificaciones, revelando un incremento en el uso de técnicas de aprendizaje automático (como redes neuronales y máquinas de vectores de soporte). Los modelos de predicción se aplican principalmente a edificios residenciales, comerciales y educativos. Además, el estudio identifica técnicas de predicción como ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal ARIMA) y SARIMAX (Seasonal ARIMA with Exogenous Variables), describiendo sus principios teóricos, etapas generales de construcción, métricas de evaluación, y las fortalezas y debilidades de cada técnica. También presenta criterios (recursos disponibles, habilidades del equipo, variables ambientales, entre otros) para facilitar la selección de técnicas y métricas de evaluación (precisión, rendimiento y costos) en función de las características del caso de estudio.

Además, con el estudio se identifican métodos para predecir el consumo de energía en edificaciones, explicando sus fundamentos teóricos, fases de creación de un modelo predictivo, y métricas de evaluación. Analiza las ventajas y desventajas de cada método y presenta pautas para seleccionar un método de predicción y métricas de evaluación según las características del caso de estudio. Destaca una tendencia hacia el uso de técnicas de aprendizaje automático y la aplicación predominante de modelos de predicción de consumo energético en edificios residenciales, comerciales y educativos.

En la investigación se concluye que la predicción del consumo energético en edificios es esencial para la planificación, gestión y conservación de la energía. La diversidad de técnicas de predicción puede representar un desafío para investigadores e ingenieros interesados en el comportamiento energético de las

edificaciones. No obstante, la aplicación de técnicas de aprendizaje automático y modelos de predicción de consumo energético resulta efectiva para comprender la eficiencia energética de los edificios. La selección adecuada de una técnica de predicción y métricas de evaluación es crucial para la toma de decisiones informadas sobre eficiencia energética en edificaciones.

El estudio presentado en [49] analiza la relación matemática entre el consumo de energía y sus variables influyentes mediante un modelo estadístico para predecir el consumo de energía eléctrica en la Universidad de Cienfuegos. El modelo permite prever el consumo futuro de energía eléctrica en la universidad a través de varias etapas, que incluyen análisis de componentes principales, regresión lineal y diseño de experimentos sobre factores como becas, computadoras, aires acondicionados, matrícula total, estudiantes de curso regular diurno, y trabajadores diurnos y nocturnos, junto con el consumo energético.

El análisis de variables de este estudio permite identificar un modelo para la predicción futura de consumo energético. Las variables analizadas son: consumo energético asociado a becados, computadoras, aires acondicionados, matrícula total, estudiantes de curso regular diurno, y trabajadores diurnos y nocturnos.

Las conclusiones indican que los hallazgos sobre el comportamiento energético de los factores analizados contribuyen a comprender mejor las variables en estudio y a seleccionar un modelo adecuado para prever el consumo energético futuro en la Universidad de Cienfuegos, destacando la importancia de la eficiencia y el ahorro energético. El estudio examina factores clave como la carga térmica, el nivel de ocupación y las condiciones climáticas, lo que permite una comprensión más detallada del comportamiento energético de las edificaciones en estudio.

El estudio presentado en [50] propone un método basado en la recopilación de datos reales para estimar de manera eficiente la eficiencia energética y la calidad del aire en hogares, considerando variables como hábitos de los residentes, normativas y tipos de construcción. El método emplea un modelo estadístico del comportamiento térmico de un edificio fundamentado en el principio de conservación de la energía. Se distinguen dos categorías de modelos: los de parámetros distribuidos y los de parámetros concentrados. Los primeros determinan la temperatura local en cada punto del entorno tridimensional, mientras que los segundos lo hacen en un número limitado de puntos o proporcionan una temperatura promedio en ciertos subconjuntos del medio, como las habitaciones de un edificio.

Además, este método emplea modelos que estiman la renovación del aire mediante la medición de componentes como un gas trazador o un contaminante común, como el CO_2. Estos modelos se basan en balances de masa que permiten

determinar continuamente las condiciones de humedad y la concentración de CO_2 en cada área. Dado el carácter aleatorio de ciertos factores relacionados con las actividades y hábitos de los ocupantes, se realiza un análisis estadístico de los datos utilizando un modelo de regresión lineal para estimar el impacto de las acciones de rehabilitación energética en el consumo de energía en edificios residenciales, lo que permite evaluar la efectividad de distintas estrategias de mejora en la eficiencia energética.

La recopilación de datos experimentales se lleva a cabo mediante el uso de sensores de calidad del aire y ambiente térmico PCE-GA 70, los cuales pueden medir la concentración de CO_2, la temperatura ambiente y la humedad relativa del aire. Los datos de las condiciones climáticas externas fueron proporcionados por la Conselleria de Medio Ambiente, Territorio e Infraestructuras de la Xunta de Galicia, tomados en sus estaciones meteorológicas ubicadas en Ferrol, España. Se utiliza el software comercial EES (Engineering Equation Solver, para sistemas operativos Microsoft Windows) para realizar los cálculos necesarios, para estimar el impacto de las acciones de rehabilitación energética en el consumo de energía de edificios residenciales.

Se concluye que durante los meses de invierno, primavera y otoño se evidencia un ahorro significativo en calefacción, mientras que el balance energético es similar en verano. La aplicación del método propuesto demuestra su eficacia para evaluar el impacto en el ahorro energético de las acciones de rehabilitación realizadas previamente, facilitando la evaluación de las labores llevadas a cabo. Además, el método descrito representa una alternativa rápida y eficaz para la planificación de proyectos de rehabilitación de la eficiencia energética, con la particularidad de estar basada en la toma de datos reales en puntos estratégicos del propio edificio. También se demostra la necesidad de tener en cuenta más variables en la construcción de los edificios para conseguir una óptima habitabilidad.

La investigación presentada en [51] analiza datos de consumo energético en edificios públicos con el objetivo de identificar tendencias, patrones y factores que influyen en el consumo de energía en estos edificios públicos. Metodológicamente aplica técnicas de análisis de datos, como el análisis de series temporales como Modelos ARIMA, SARIMA y SARIMAX, y la modelización predictiva, Redes Fuzzy, Sistemas Expertos, Modelos de Mínimos Cuadrados Parciales y Modelos de Regresión Lineal.

Se pretende elaborar modelos precisos que posibiliten predecir el consumo energético futuro. Este enfoque tiene la finalidad de optimizar el uso de la energía, mejorar la eficiencia energética y disminuir los costos asociados al consumo de energía en edificios públicos.

En la tesis, se desarrolla una metodologías integral, que combinaba técnicas de minería de datos, análisis de perfiles de consumo y modelización predictiva para comprender y anticipar el comportamiento energético de los edificios públicos analizados, y para la visualización de patrones de consumo energético y su predicción, la representación gráfica de los perfiles de consumo de energía de diferentes edificios que permiten su comparación y agrupación, lo que facilita la comprensión de su distribución y la identificación de aquellos con patrones de consumo similares. Esta información es útil para la implementación de estrategias de ahorro y eficiencia energética.

Como conclusión, se perfecciona el sistema de vigilancia mediante una arquitectura de tres niveles, la cual simplifica la adquisición, almacenamiento y visualización de los datos. Esta estructura fue concebida de forma modular y escalable, con una capa intermedia que ejecuta tareas específicas, pero utiliza el mismo hardware. Esta estructura se implementa con datos eléctricos de medidores instalados en los edificios, pero de fácil adaptación para trabajar con otros tipos de datos.

El estudio presentado en [52] aborda la elaboración y verificación de un modelo de predicción de energía para el hotel Meliá Habana en La Habana, Cuba. El modelo utiliza el método de series de tiempo radiantes para estimar la carga térmica de las áreas residenciales del hotel y se desarrolla mediante `MatLab` [53].

El método de series de tiempo radiantes se emplea para calcular la carga térmica de las zonas habitacionales de la instalación. La validación del modelo se realiza experimentalmente, mediante mediciones reales del consumo energético diario del hotel. La utilización de datos reales en la validación es fundamental para garantizar la precisión y eficacia del modelo, lo que lo convirtie en una herramienta útil para el análisis del comportamiento energético y la implementación de estrategias avanzadas de control en el hotel Meliá Habana.

Se determina que el modelo obtenido confirmo que la carga térmica de cada habitación del hotel, está influenciada por su ubicación geográfica dentro de la instalación hotelera y por sus características constructivas.

En conclusión, la revisión de antecedentes revela un avance en las metodologías de predicción de consumo energético, con una tendencia hacia la integración de técnicas de aprendizaje automático y modelización estadística en edificios residenciales, comerciales y públicos. Estos estudios destacan la importancia de adaptar las técnicas de modelización a las características específicas del entorno y los usos de cada edificación, evidenciando su potencial para contribuir al ahorro energético y la sostenibilidad en el sector de la construcción.

6 Metodología

Dado el número limitado de observaciones disponibles (49 proyectos) y la presencia de clases desbalanceadas tanto en la variable de certificación (certificado/no certificado) como en los niveles de ahorro energético (bajo, medio y alto), este estudio propone una metodología diseñada para enfrentar dos desafíos principales: el alto número de variables en relación con el tamaño muestral y el desequilibrio entre clases en los métodos de clasificación.

La estrategia metodológica combina técnicas de preprocesamiento, reducción de dimensionalidad, modelado supervisado y validación rigurosa, con el fin de garantizar un análisis confiable y generalizable. La metodología del estudio se estructura en cinco etapas consecutivas:

1. **Preprocesamiento de Datos**

   Esta etapa se centra en el análisis exploratorio y la preparación del conjunto de datos para su uso en los modelos posteriores. Se examina la distribución del consumo energético y su relación con las características estructurales y operativas de los edificios incluidos en la muestra. Se realiza una revisión detallada de valores faltantes, la detección y análisis de valores atípicos, así como un análisis descriptivo multivariado de las variables predictoras y de respuesta.

   Para las variables cuantitativas explicativas se aplican transformaciones de Box-Cox y Yeo-Johnson, según corresponda, con el fin de aproximar la normalidad. Posteriormente, se realiza un escalado robusto mediante el rango intercuartílico (IQR). Se generan también las variables categóricas necesarias para los modelos de clasificación: la variable de certificación energética (certificado/no certificado) y el nivel de ahorro energético, categorizado en tres niveles (bajo, medio y alto) según terciles del porcentaje de ahorro.

Se evalúan las asociaciones entre variables cuantitativas mediante el coeficiente de correlación de Kendall, dada la presencia de asimetrías y sesgos en la distribución. Con base en este análisis se seleccionan las variables más relevantes para los modelos posteriores. Finalmente, se crean variables dummy para representar las variables categóricas, eliminando la categoría de referencia y aplicando la restricción de suma cero en los modelos correspondientes para garantizar la identificabilidad.

2. **Reducción de Dimensionalidad Mediante PCA Robusto**

   En esta etapa se implementa un Análisis de Componentes Principales (PCA) robusto usando las variables cuantitativas transformadas, con el objetivo de reducir la dimensionalidad del conjunto de variables explicativas mientras se mitiga la influencia de valores atípicos. Se analizan las gráficas de dispersión de las componentes principales para explorar posibles patrones de agrupación entre los proyectos, tanto en función del estado de certificación (certificado/no certificado) como de los niveles de ahorro energético (bajo, medio y alto).

   Se seleccionan aquellas componentes que, en conjunto, explican al menos el 60 % de la varianza total, las cuales se utilizan posteriormente como predictores en los modelos de regresión beta y Análisis Discriminante Robusto (RDA). Asimismo, se identifican las variables originales con mayor contribución a cada componente seleccionada.

3. **Modelado Mediante Regresión Beta**

   Esta etapa abarca la implementación y comparación de distintos modelos de regresión beta para predecir el porcentaje de ahorro energético. Dado que la variable respuesta presenta valores iguales a cero, se aplica la transformación recomendada por Smithson y Verkuilen para adecuarla al dominio $(0, 1)$. Se evalúan tres escenarios de modelado:

   a) Usando únicamente las componentes principales obtenidas del PCA robusto.

   b) Combinando las componentes principales con variables categóricas transformadas en variables dummy (sin categoría de referencia).

   c) Utilizando directamente las variables cuantitativas transformadas (sin PCA) junto con las variables dummy derivadas de las variables categóricas, también sin categoría de referencia.

   Para cada escenario se aplica validación cruzada 5-Fold utilizando los 44 proyectos (4 no certificados) de la base de entrenamiento. El modelo resul-

tante se evalúa posteriormente en un conjunto de prueba independiente compuesto por 5 proyectos (1 no certificado). Como métrica de desempeño se utiliza el Error Absoluto Medio (MAE). Se comparan los modelos ajustados con diferentes funciones de enlace (logit, probit y complemento–log–log), considerando además las métricas AIC, BIC, pseudo-R^2 y log-verosimilitud. Estos análisis incluyen la validación de supuestos y la comparación del ajuste global de los modelos.

Se selecciona el modelo beta que conserva únicamente los predictores estadísticamente significativos, cumple los supuestos distribucionales y presenta el MAE más bajo; dentro de este modelo, la importancia de cada variable se determina por la magnitud absoluta de su coeficiente, indicativa de la fuerza de su impacto sobre el porcentaje de ahorro, y por la coherencia operativa de dicho impacto, es decir, que su signo y tamaño sean razonables y compatibles con la realidad técnica y la gestión energética.

4. **Clasificación mediante Análisis Discriminante Robusto (RDA)**

   Se integran las variables dummies de todas las categóricas (sin categoría de referencia) y las componentes seleccionadas mediante PCA robusto. Se añaden dos etiquetas de salida: la binaria *Certificación (Sí / No)* y la triclase *Nivel de ahorro (Bajo / Medio / Alto)*.

   Se evalúan dos escenarios de modelado:

   a) Usando únicamente las componentes principales obtenidas del PCA robusto.

   b) Combinando las componentes principales con variables categóricas transformadas en variables dummy (sin categoría de referencia).

   Respecto a la propuesta de la metodología se tienen las siguientes consideraciones incluyendo posibles relaciones de costos.

   a) *Caso binario*: En este escenario binario (certificación *sí/no*) los costos de error son marcadamente asimétricos: un falso negativo (FN) supone perder incentivos, mientras que un falso positivo (FP) sólo genera una auditoría innecesaria. Suponiendo que los aciertos (VP y VN) resultan económicamente neutros,[1] la razón de costos se puede aproximar en la realidad por

$$\frac{C_{\text{FN}}}{C_{\text{FP}}} \approx 5{:}1.$$

[1] Se considera que una certificación válida o un rechazo correcto no conllevan costos adicionales relevantes.

Para reflejar esta prioridad se optimiza la métrica de la familia F_β, dada matemáticamente como:

$$F_\beta \;=\; \frac{(1+\beta^2)\,\mathrm{Prec}\,\mathrm{Sens}}{\beta^2\,\mathrm{Prec}+\mathrm{Sens}},$$

donde $\mathrm{Sens} = \mathrm{VP}/(\mathrm{VP}+\mathrm{FN})$ y $\mathrm{Prec} = \mathrm{VP}/(\mathrm{VP}+\mathrm{FP})$. El parámetro β^2 actúa como *peso relativo* de la *Sensibilidad* (recuperar proyectos certificables) sobre la *Precisión* (evitar auditorías innecesarias). Fijar $\beta = 2$, esto es, $\beta^2 = 4$, otorga a la *Sensibilidad* un peso cuatro veces mayor que a la *Precisión*, aproximando la razón de costos $5{:}1$ sin descuidar del todo los falsos positivos. El valor es compatible con los requisitos operativos $\mathrm{Sens} \geq 0{,}90$ y $\mathrm{Spec} \geq 0{,}80$, que ya privilegian la detección de proyectos realmente certificables.

Para evitar sobre penalizar los FP se adopta una versión *más moderada* $2{:}1$. El análisis económico actual fija $C_{\mathrm{FN}} : C_{\mathrm{FP}} \simeq 2{:}1$; por ello se elige

$$\beta^2 = 2 \;(\beta = \sqrt{2}),$$

de modo que la *Sensibilidad* pesa el doble que la *Precisión*, en línea con el impacto financiero de cada tipo de error.

Para reflejar la razón económica $C_{\mathrm{FN}} : C_{\mathrm{FP}} \simeq 2{:}1$ se adopta la siguiente *matriz de costos* $\mathbf{C}$ (columnas = clase real, filas = predicción):

$$\mathbf{C} \;=\; \begin{array}{c|cc} & y=0 & y=1 \\ \hline \hat{y}=0 & 0 & 2 \\ \hat{y}=1 & 1 & 0 \end{array}$$

- $C_{00} = 0$: *verdadero negativo* (VN) — decisión correcta, sin costo.
- $C_{01} = 2$: *falso negativo* (FN) — se descarta un proyecto certificable; pérdida mayor.
- $C_{10} = 1$: *falso positivo* (FP) — auditoría innecesaria; costo menor.
- $C_{11} = 0$: *verdadero positivo* (VP) — decisión correcta, sin costo.

Esta tabla guía:

- la selección del umbral óptimo t^* minimizando $C(t) = 2\,\mathrm{FN}(t) + 1\,\mathrm{FP}(t)$ con $\mathrm{Spec} \geq 0{,}50$;
- la ponderación de la métrica $F_{\sqrt{2}}$, donde $\beta^2 = 2$ concede el doble de peso a la *Sensibilidad*;
- los pesos de clase $2{:}1$ aplicados en el entrenamiento del RDA.

El ajuste de los hiperparámetros γ y λ del RDA binario (*certificado/no certificado*) sigue una búsqueda sobre una cuadrícula de $28 \times 28 = 784$

combinaciones, uniformemente espaciadas en el intervalo $[0{,}10,\ 0{,}91]$ con paso fijo $0{,}03$. Cada par (γ, λ) se evalúa mediante *validación cruzada repetida*: 4 pliegues estratificados que se recrean 10 veces con particiones aleatorias (40 re-muestreos en total). Se usan pesos de clase 2:1 dentro de cada fold. En cada fold se calculan cinco métricas clave: *Sensibilidad, Especificidad, Balanced Accuracy, AUC-PR* y $F_{\sqrt{2}}$.

Los 40 valores de cada métrica se promedian, produciendo un resultado para cada (γ, λ).

Después de promediar las métricas por combinación se aplica un filtro jerárquico de umbrales:

$$\text{Sensibilidad} \geq \{0{,}90; 0{,}85; 0{,}80\}, \qquad \text{Especificidad} \geq 0{,}80 \rightarrow 0{,}50.$$

Dentro del primer nivel que cumple los dos umbrales se elige el modelo de *mayor* F_2 (desempate por *Balanced Accuracy*); si ninguno los supera, se toma el F_2 máximo global.

Con el fin de determinar el umbral de decisión t^*, para cada uno de los 40 folds se minimiza $C(t) = 2\,\text{FN}(t) + 1\,\text{FP}(t)$ imponiendo Spec $\geq 0{,}50$; t^* se fija como la *mediana* de los 40 valores t_i^*.

Se realiza evaluación externa con LOOCV. Los hiperparámetros ganadores (γ^*, λ^*) se emplean para re entrenar el modelo con los 49 proyectos y los pesos 2:1. En cada iteración se predice el caso excluido, se aplica el corte global t^* y se registran *Sensibilidad, Especificidad, Balanced Accuracy, AUC-PR* y $F_{\sqrt{2}}$.

El diagnóstico se realiza con la matriz de confusión con el total de aciertos y errores, curvas ROC y PR a partir de las probabilidades LOOCV; se reportan los AUC correspondientes.

La coherencia métrica–costo se garantiza porque $F_{\sqrt{2}}$, el umbral t^* y los pesos 2:1 reflejan la misma prioridad económica (FN $\gg$ FP).

La doble validación 4×10 CV para la búsqueda interna + LOOCV externa evita sobre-ajuste y proporciona estimaciones casi insesgadas de desempeño fuera de muestra.

b) *Caso triclases*: Se definen tres clases equiespaciadas por terciles: *Bajo*, *Medio* y *Alto*. Los tamaños son casi idénticos, de modo que el desbalance deja de ser crítico; sin embargo, *Medio* y *Alto* son los niveles de mayor relevancia práctica al corresponder con proyectos potencialmente certificables.

El impacto económico es *asimétrico*: *sub-estimar* el nivel real (p. ej. predecir *Bajo* cuando el proyecto es *Alto*) implica perder incentivos y se

penaliza el doble, mientras que *sobre-estimar* genera sólo una revisión extra y cuesta la mitad. Esto se traduce en la matriz de costes $\mathbf{C}$ (filas = predicción $\hat{c}$, columnas = clase real c):

$$\mathbf{C} = \begin{array}{c|ccc} & c=1\ (B) & c=2\ (M) & c=3\ (A) \\ \hline \hat{c}=1\ (B) & 0 & 2 & 2 \\ \hat{c}=2\ (M) & 1 & 0 & 2 \\ \hat{c}=3\ (A) & 1 & 1 & 0 \end{array}$$

- Diagonal $C_{cc} = 0$: aciertos sin costo.
- $C_{\text{infra}} = 2$: el modelo predice un nivel más bajo que el real (p.ej. predice "Bajo" cuando el nivel verdadero es "Medio" o "Alto"), ocasionan pérdida de incentivos.
- $C_{\text{sobre}} = 1$: el modelo predice un nivel más alto que el real (p.ej. predice "Alto" cuando el nivel verdadero es "Medio" o "Bajo"), sólo provocan auditorías adicionales.

Esta tabla orienta a:

- la búsqueda del umbral global t^*, minimizando $C(t) = 2\,\text{FN}(t) + 1\,\text{FP}(t)$ bajo la restricción $\text{Spec}_{\text{macro}}(t) \geq 0{,}50$;
- la elección del $F^{\text{macro}}_{\sqrt{2}}$ como métrica objetivo (doble peso a la Sensibilidad de cada clase);
- la penalización del algoritmo en la fase de entrenamiento, aplicando un esquema de costos coherente con la matriz anterior.

Como métrica principal se optimiza el *macro-*$F_{\sqrt{2}}$:

$$F^{\text{macro}}_{\sqrt{2}} = \frac{1}{3} \sum_{c \in \{\text{Bajo, Medio, Alto}\}} \frac{(1+\sqrt{2}^2)\,\text{Prec}_c\,\text{Sens}_c}{\sqrt{2}^2\,\text{Prec}_c + \text{Sens}_c},$$

con $\beta = \sqrt{2}$ para otorgar a la sensibilidad el *doble* de peso que a la precisión, en coherencia con el costo 2 : 1.

El procedimiento de calibración de hiperparámetros conlleva a considerar:

- *Grilla.* $28 \times 28 = 784$ valores $\gamma, \lambda \in [0{,}10, 0{,}91]$ (paso $0{,}03$).
- *Validación cruzada interna.* 4-fold estratificada, 10 repeticiones (40 re-muestreos). No se aplican pesos ni SMOTE por la buena distribución de clases. En cada fold se registran $F^{\text{macro}}_{\sqrt{2}}$, $\text{Sens}_{\text{macro}}$, $\text{Spec}_{\text{macro}}$, $\text{BalAcc}_{\text{macro}}$, y se promedian sobre los 40 re-muestreos.
- *Filtro jerárquico.* $\text{Sens}_{\text{macro}} \geq \{0{,}90, 0{,}85, 0{,}80\}$ y $\text{Spec}_{\text{macro}} \geq 0{,}80 \rightarrow 0{,}50$. En el primer nivel que cumple ambos cortes se elige la pareja

(γ, λ) con mayor $F_{\sqrt{2}}^{\text{macro}}$ (desempate por $\text{BalAcc}_{\text{macro}}$); si ningún modelo pasa el filtro, se toma el máximo global.

- *Umbral global* t^*. Para cada uno de los 40 folds se minimiza $C(t) = 2\,\text{FN}(t) + 1\,\text{FP}(t)$ bajo la restricción $\text{Spec}_{\text{macro}}(t) \geq 0{,}50$. La mediana de los 40 valores t_i^* define el corte definitivo t^*.

Puesto que el $n = 49$ es pequeño, se realiza un LOOCV externo. Para tal efecto se re-entrena con los 49 proyectos y, en cada iteración, se deja fuera exactamente uno; sobre las probabilidades resultantes se aplica el mismo t^* y se generan las curvas ROC/PR *one-vs-rest* y las métricas macro-promediadas.

Posteriormente, en la etapa de diagnóstico final, se reportan:

- matriz de confusión global,
- curva ROC–macro y PR–macro,
- valores finales de $F_{\sqrt{2}}^{\text{macro}}$, $\text{Sens}_{\text{macro}}$, $\text{Spec}_{\text{macro}}$ y $\text{BalAcc}_{\text{macro}}$, todos calculados con el corte t^*.

Por cada uno de los escenarios se determinan hiperparámetros óptimos, umbral óptimo. Además, se realiza el entrenamiento y evaluación del modelo, y se determinan variables más importantes.

A continuación se explica la metodología por pasos, para el caso binaria ($k = 2$):

a) *Selección simultánea de* (γ, λ) *y del umbral* t^* mediante validación cruzada estratificada *4-fold* $\times$ *10 réplicas* (40 re-muestreos):

1) Se recorre una rejilla relativamente fina $\gamma, \lambda \in [0{,}10, 0{,}91]$ con paso $0{,}03$.
2) En cada re-muestreo se calculan *Sensibilidad*, *Especificidad*, BalAcc y $F_{\sqrt{2}}$ (*recall* con doble peso). Los pesos de clase en la pérdida logarítmica interna son $2{:}1$ (FN cuesta el doble que FP).
3) *Filtro jerárquico*: se exige primero $\text{Sens} \geq \{0{,}90, 0{,}85, 0{,}80\}$ y, dentro de cada escalón, $\text{Spec} \geq 0{,}80 \to 0{,}50$.
4) Entre los modelos que superan el filtro se escoge (γ^*, λ^*) con F_2 máximo; en caso de empate se prefiere mayor BalAcc.
5) Con las predicciones de ese par se barre $t \in [0, 1]$ cada $0{,}001$. Se retienen los umbrales que satisfacen $\text{Spec}(t) \geq 0{,}50$ y, entre ellos, se minimiza el coste $C(t) = 2\,\text{FN}(t) + 1\,\text{FP}(t)$. La mediana de los 40 valores obtenidos es el punto-corte global t^*.

b) *Evaluación externa con LOOCV* (49 pliegues *leave-one-out*):

1) Se fija (γ^*, λ^*) y se re-entrena el modelo en cada iteración dejando fuera un único proyecto; los pesos 2:1 se mantienen.
2) Las probabilidades de los 49 proyectos fuera-de-muestra se cortan con t^* para obtener la etiqueta final.
3) Se calculan *Sensibilidad*, *Especificidad*, BalAcc y $F_{\sqrt{2}}$ globales; además se grafican la matriz de confusión, la curva ROC (con AUC) y la curva Precision–Recall (con $\mathrm{AUC}_{\mathrm{PR}}$).

c) *Interpretabilidad*: con `varImp()` se extrae la importancia de los predictores (las componentes $\mathrm{PC}_1 - \mathrm{PC}_4$ y las dummies). Se grafica un bar-plot horizontal ordenado por la columna *Overall*.

Los resultados se enlazan de la siguiente manera: (γ^*, λ^*) se utilizan dos veces, para calcular t^* en la fase de búsqueda y como hiper-parámetros fijos en el LOOCV. El umbral t^* sólo se aplica *después* del LOOCV para transformar probabilidades en clases, de modo coherente con el coste $2\,\mathrm{FN} + 1\,\mathrm{FP}$. Los 49 proyectos actúan cada uno como *test* exactamente una vez; ya no queda un conjunto externo sin evaluar.

Para el caso triclases ($k = 3$), el procedimiento anterior se reutiliza casi por completo; sólo cambian (i) la métrica–objetivo, (ii) el criterio de corte de probabilidad y (iii) la forma de ponderar los errores. A continuación se destacan únicamente los pasos que difieren del caso binario:

a) *Métrica de optimización.* Se reemplaza $F_{\sqrt{2}}$ por su versión *macro-promediada*[2] El factor $\beta^2 = 2$ mantiene la razón coste–beneficio $C_{\mathrm{FN}}{:}C_{\mathrm{FP}} \simeq 2{:}1$.

b) *Ponderación de la pérdida interna.* Las tres clases están casi balanceadas, por lo que *no se utilizan pesos de clase* en la función de coste logarítmico del entrenamiento.

c) *Filtro jerárquico.* Se sustituyen Sens y Spec por sus promedios macro:

$$\mathrm{Sens}_{\mathrm{macro}} \geq \{0{,}90,\ 0{,}85,\ 0{,}80\}, \qquad \mathrm{Spec}_{\mathrm{macro}} \geq 0{,}80 \rightarrow 0{,}50.$$

La selección final dentro del primer nivel factible se hace por $F_{\sqrt{2}}^{\mathrm{macro}}$ (desempate por $\mathrm{BalAcc}_{\mathrm{macro}}$).

d) *Umbral global t^*.* Aunque el problema es multiclase, al negocio sólo le interesa saber si el nivel asignado por el modelo *se acepta o se rectifica*. Se define por tanto un coste binario $C(t) = 2\,\mathrm{FN}(t) + 1\,\mathrm{FP}(t)$ sobre la

[2] Cada clase se enfrenta a las otras dos (*one–vs–rest*).

clase predicha (ganadora del argmax de probabilidades) respetando $\mathrm{Spec}_{\mathrm{macro}}(t) \geq 0{,}50$; el procedimiento de barrido y mediana de los 40 valores t_i^* es idéntico al caso $k = 2$.

e) *Evaluación externa.* Por defecto se reserva el *hold-out* de 5 proyectos exclusivamente para esta etapa, evitando LOOCV.

f) *Interpretabilidad.* Los pasos restantes —interpretabilidad con `varImp()`, búsqueda sobre la cuadrícula de 784 combinaciones y reutilización coherente de $(\gamma^*, \lambda^*, t^*)$— siguen exactamente el mismo esquema descrito para el problema binario.

5. **Predicciones**. Finalmente, considerando las mejores configuraciones de los modelos, se analizan las predicciones obtenidas tanto para el porcentaje de ahorro energético (regresión beta) como para la clasificación del nivel de ahorro y la certificación (RDA).

La implementación de la metodología en este estudio se realiza utilizando el software estadístico `R` [54]. Para el análisis de componentes principales robusto (PCA robusto) se emplea la función `PcaHubert` del paquete `rrcov` [55]. La transformación Box-Cox se aplica mediante las funciones `powerTransform` y `bcPower` de la biblioteca `car` [56]. El escalado robusto de las variables se realiza utilizando la función `mad` en conjunto con `scale` de la librería `stats` [57]. La regresión beta se implementa mediante la función `betareg` del paquete `betareg`[58]. Para el análisis discriminante robusto se utiliza la función `Linda` del paquete `rrcov` [55]. Finalmente, la validación cruzada se lleva a cabo con las funciones `trainControl` y `train` de los paquetes `caret` [59] y `klaR`[60], respectivamente.

7 Datos

Los datos utilizados en este estudio comprenden 3 variables relacionadas con el consumo energético, 27 características estructurales y operativas de 49 edificaciones ubicadas en Colombia, 44 con certificación de ahorro de energía y 5 no certificados. Estas características incluyen 23 variables cuantitativas y 4 cualitativas, además de la estimación del consumo y ahorro energético correspondiente a cada edificación.

Las variables operativas y estructurales corresponden a aquellas empleadas por la metodología *ASHRAE* para estimar el consumo y el ahorro de energía. Los consumos y porcentajes de ahorro utilizados en este estudio no corresponden a registros reales, sino a estimaciones generadas mediante la metodología *ASHRAE*. En este trabajo, dichas variables, estimadas por los modelos físicos de la metodología, son tratadas como variables aleatorias.

Desde un punto de vista técnico, *EnergyPlus* y *DOE-2*, utilizadas en la metodología *ASHRAE*, son herramientas deterministas, ya que producen resultados consistentes al emplear los mismos parámetros de entrada para una edificación. No obstante, debido a la incertidumbre inherente en los datos de entrada, los supuestos simplificadores que subyacen a las simulaciones y las variaciones observadas entre escenarios prácticos, en este trabajo se opta por tratar los resultados de las simulaciones como variables aleatorias en los análisis estadísticos. Este enfoque permite capturar la variabilidad asociada al proceso de simulación y facilita la construcción de modelos predictivos más representativos de la realidad.

En la Tabla 1 se representa una lista detallada de variables, su notación y una breve descripción: *Tipología* (TIP), *Ciudad* (CIU), *Zona climática* (ZCL), *Certificación* (CER), *Área total* (ART), *Área acondicionada* (AAC), *Porcentaje WWR* (WWR), *Área de cubierta* (ACB), *Altura* (ALT), *Lado más ancho* (LMA), *Esbeltez* (ESB), *Orientación respecto al norte* (ORN), *Área de fachada Oriente* (AFO), *Área de fachada Occidente*

(AFOC), *Área de ventanería Oriente* (AVO), *Área de ventanería Occidente* (AVOC), *U-Value muros exteriores* (UME), *U-Value muros cubierta* (UMC), *Reflectancia muros* (RFM), *Reflectancia cubiertas* (RFC), *U-Value vidrio* (UVG), *SHGC Vidrio* (SHV), *VLT Vidrio* (VLT), *Aire acondicionado* (AIA), *Consumo Propuesto* (CPR), *Consumo Base* (CBA), *Energía renovable* (ERN), *EUI Propuesto* (EUP), *EUI Base* (EUB), *Porcentaje de ahorro* (PAH).

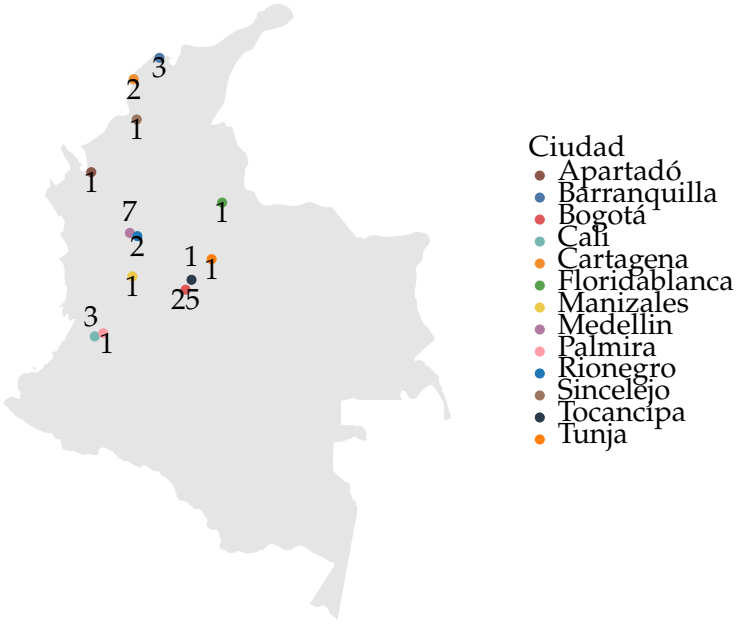

Figura 1: Distribución espacial del número de proyectos por ciudad en Colombia.

La *tipología* de un proyecto se refiere a su uso principal, determinado por el objetivo de su construcción. Por ejemplo, si el proyecto se destina a las actividades administrativas y operativas de una empresa, su tipología se clasifica como *oficinas*. Esta variable puede adoptar diferentes valores cualitativos según el propósito del proyecto, tales como: *biblioteca*, *salud*, *hotel*, *oficinas*, *residencial*, *comercial* o *universidad*.

La *ciudad* se refiere a la ubicación geográfica del edificio. Las ciudades consideradas en este estudio son: *Barranquilla*, *Bogotá*, *Cali*, *Cartagena*, *Floridablanca*, *Manizales*, *Medellín*, *Palmira*, *Quito* (Ecuador, una ciudad con características climáticas similares a las de Colombia), *Rionegro*, *Sincelejo*, *Tocancipá* y *Tunja*. En la Figura 1 se puede observar la distribución espacial del número de proyectos por ciudad en Colombia.

Tabla 1: Características principales del consumo energético y los aspectos estructurales y operativos de los edificios certificados en la muestra.

No.	Nombre de la variable	Notación	Descripción
1	Tipología	TIP	Uso principal del proyecto (ej.: oficinas). Valores cualitativos según el propósito de construcción.
2	Ciudad	CIU	Ciudad donde está el edificio. Ejemplo: Barranquilla, Bogotá, Cali, etc.
3	Zona climática	ZCL	Valor según *ASHRAE*: número para temperatura (0 a 8) y letra para humedad (A, B o C). Ej.: 1A.
4	Certificación	CER	Indica si el proyecto cuenta con certificación en sostenibilidad según *ASHRAE*.
5	Área total	ART	Área total del proyecto en metros cuadrados (m^2).
6	Área acondicionada	AAC	Área destinada para el sistema HVAC en metros cuadrados (m^2).
7	Porcentaje WWR	WWR	Relación ventana-pared del proyecto (0 a 1).
8	Área de cubierta	ACB	Área de la cubierta en metros cuadrados (m^2).
9	Altura	ALT	Altura del edificio en metros (m).
10	Lado más ancho	LMA	Longitud del lado más largo del proyecto en metros (m).
11	Esbeltez	ESB	Relación ancho/alto. Edificio esbelto si es mayor o igual a 0.1.
12	Orientación respecto al norte	ORN	Ángulo respecto al norte en grados, entre 0^o y 360^o.
13	Área de fachada Oriente	AFO	Área orientada al oriente en metros cuadrados (m^2).
14	Área de fachada Occidente	AFOC	Área orientada al occidente en metros cuadrados (m^2).
15	Área de ventanería Oriente	AVO	Área de ventanería orientada al oriente en metros cuadrados (m^2).
16	Área de ventanería Occidente	AVOC	Área de ventanería orientada al occidente en metros cuadrados (m^2).
17	U-Value muros exteriores	UME	Coef. de transmitancia térmica de muros exteriores (W/m^2-K).
18	U-Value cubierta	UMC	Coef. de transmitancia térmica de la cubierta (W/m^2-K).
19	Reflectancia muros	RFM	Capacidad de los muros para reflejar radiación solar (0 a 1).
20	Reflectancia cubiertas	RFC	Capacidad de la cubierta para reflejar radiación solar (0 a 1).
21	U-Value vidrio	UVG	Coef. de transmitancia térmica del vidrio (W/m^2-K).
22	SHGC Vidrio	SHV	Solar Heat Gain Coefficient: radiación transmitida/total en vidrio (0 a 1).
23	VLT Vidrio	VLT	Visible Light Transmittance: luz natural que permite el vidrio (0 a 1).
24	Aire acondicionado	AIA	Tipo de sistema de aire acondicionado (ej.: VRF, chiller). Incluye ventilación natural.
25	Consumo Propuesto	CPR	Consumo de energía anual del proyecto en kWh.
26	Consumo Base	CBA	Consumo anual en línea base *ASHRAE* en kWh.
27	Energía renovable	ERN	kWh generados anualmente por fuentes renovables, como paneles fotovoltaicos.
28	EUI Propuesto	EUP	Relación consumo propuesto/área (kWh-año/m^2).
29	EUI Base	EUB	Relación consumo en línea base/área (kWh-año/m^2).
30	Porcentaje de ahorro	PAH	Porcentaje de ahorro estimado según *ASHRAE*.

La *certificación* se refiere al tipo de certificación en sostenibilidad que recibió el proyecto. Esto para asegurar que se sigan los lineamientos necesarios de *ASHRAE*. Las certificaciones incluyen: *LEED 2009 ID+C: Commercial interiors*, *LEED*

v2009 C&S Development, *LEED V2009: NC*, *LEED V2009: Retail*, *LEED v4 BD+C: Core and Shell*, *LEED v4 BD+C: Healthcare*, *LEED v4 BD+C: NC*, y *LEED v4 BD+C: Retail*.

La *zona climática* es un valor cualitativo asignado a la ciudad o ubicación del proyecto según los lineamientos de ASHRAE, compuesto por un número y una letra. Los números van del 0 al 8, donde 0 indica un clima extremadamente cálido y 8 un clima extremadamente frío. Las letras, que van de A a C, indican el nivel de humedad: A para húmedo, B para seco y C para marino. Este valor puede variar según el criterio del profesional encargado de la modelación.

- *Húmedo (A)*: Ubicaciones que no son marinas ni secas.
- *Seco (B)*: Ubicaciones que no son marinas y cumplen con el criterio $P < 2 \times (T + 7)$, donde P es la precipitación anual en cm y T es la temperatura media anual en °C.
- *Marino (C)*: Ubicaciones que cumplen con los siguientes criterios:

  1. La temperatura promedio del mes más frío está entre -3°C y 18°C.
  2. La temperatura del mes más cálido es menor a 22°C.
  3. Al menos cuatro meses tienen una temperatura media superior a 10°C.
  4. El mes con mayor precipitación en la estación fría tiene al menos tres veces la precipitación del mes con menor precipitación en el resto del año. La estación fría es de octubre a marzo en el hemisferio norte y de abril a septiembre en el hemisferio sur.

Los valores de *zona climática* combinados definen diferentes clasificaciones, como se muestra a continuación:

- *1A* - Muy caliente y húmedo
- *1B* - Muy caliente y seco
- *2A* - Caliente y húmedo
- *2B* - Caliente y seco
- *3A* - Tibio y húmedo
- *3C* - Tibio y marino
- *4C* - Mixto y marino

8 Resultados

Los resultados de este trabajo se presentan en cinco etapas, las cuales corresponden directamente a las fases descritas en la metodología.

8.1. Etapa 1: Preprocesamiento de los Datos

La base de datos cuenta con 49 proyectos y 30 variables, entre estas, 25 son cuantitativas y 5 cualitativas. El 94 % de los proyectos no cuenta con generación de energía por fuentes alternativas.

1. *Análisis Univariado*

   En la Tabla 2 se muestran las principales estadísticas descriptivas de las variables cuantitativas; a partir de ellas se concluye lo siguiente:

   - Un proyecto carece de aire acondicionado (AAC); por lo tanto, su acondicionamiento térmico depende exclusivamente de la ventilación natural.
   - En promedio, los edificios analizados presentan una relación ventana-muro (WWR) del 37,9 %, aproximadamente. La mediana es aproximada de 36,65 %, por lo que la mitad de las construcciones supera este umbral, un valor relativamente elevado que podría atribuirse al predominio de edificaciones comerciales y de oficinas en la muestra.
   - El área construida bruta (ACB) varía entre 0 y 28 419,32 m^2; los proyectos con ACB igual a cero corresponden a aquellos ubicados en pisos intermedios de los edificios.
   - La razón de esbeltez (ESB) muestra un primer cuartil de 0,09, lo que indica que solo el 25 % de los proyectos registra valores iguales o

inferiores a ese umbral. Dado que se considera esbelto todo edificio con $ESB > 0{,}10$, el 75 % restante de la muestra se clasifica como esbelto.

Tabla 2: Estadísticos descriptivos de las variables cuantitativas

No.	Variable	Media	Desv.Est	Min	Q1	Mediana	Q3	Max	CV
1	AAC	15111,02	23805,29	0,00	1857,58	5621,50	18722,86	117705,60	1,58
2	ACB	4939,86	6714,69	0,00	978,11	2268,50	5779,60	28419,32	1,36
3	AFO	1343,35	2016,61	0,00	283,50	714,01	1426,01	10396,30	1,50
4	AFOC	1339,49	2032,85	0,00	416,05	697,58	1026,56	10653,00	1,52
5	ALT	23,92	23,79	3,20	8,00	14,77	32,00	118,00	0,99
6	ART	27734,17	49409,98	79,10	2973,00	7441,35	31596,00	236009,70	1,78
7	AVO	415,37	694,00	0,00	55,00	182,50	387,40	3281,80	1,67
8	AVOC	387,51	691,50	0,00	52,85	154,70	482,72	3795,46	1,78
9	CBA	3520670,94	6304858,84	3875,97	313734,81	890349,90	2409659,20	25102339,10	1,79
10	CPR	2776895,97	5008616,66	3168,02	252891,03	667249,16	1885871,70	19556455,05	1,80
11	ERN	5247,62	28518,20	0,00	0,00	0,00	0,00	192108,00	5,43
12	ESB	0,42	0,50	0,04	0,09	0,21	0,49	2,05	1,21
13	EUB	129,56	83,41	24,92	73,89	110,15	146,97	403,73	0,64
14	EUP	100,93	67,73	18,90	55,57	82,47	108,38	333,80	0,67
15	LMA	81,29	55,96	16,43	48,94	61,76	90,83	275,30	0,69
16	ORN	155,34	141,22	0,00	17,00	104,50	305,00	358,00	0,91
17	PAH	0,21	0,11	0,00	0,17	0,22	0,28	0,47	0,52
18	RFC	0,39	0,14	0,30	0,30	0,30	0,45	0,82	0,37
19	RFM	0,36	0,12	0,30	0,30	0,30	0,40	0,75	0,32
20	SHV	0,54	0,18	0,24	0,43	0,50	0,69	0,90	0,33
21	UMC	1,67	1,50	0,21	0,58	0,97	2,03	5,65	0,90
22	UME	1,64	0,99	0,06	0,90	1,46	2,25	4,03	0,61
23	UVG	4,31	1,64	1,10	2,91	5,05	5,71	6,35	0,38
24	VLT	0,66	0,20	0,16	0,50	0,73	0,76	0,90	0,29
25	WWR	37,90	23,06	5,10	18,68	36,65	47,65	99,80	0,61

- En promedio, los proyectos están orientados (ORN) a 155,34° con respecto al norte; la mediana es 141,22°, de modo que la mitad supera este ángulo.
- Las reflectancias de fachada (RFM) y de cubierta (RFC) son muy similares, con valores medios de 0,39 y 0,36, respectivamente. En ambos casos, la mediana se sitúa en 0,03, lo que implica que la mitad de los proyectos presentan reflectancias inferiores a ese umbral, evidenciando un nivel de reflectancia bajo.
- Los coeficientes de transmisión térmica medios del muro exterior (UME) y de la cubierta (UMC) son 1,64 y 1,67 $W/m^2 K$, respectivamente. Las medianas, de 1,46 y 0,97 $W/m^2 K$, indican que la mitad de los proyectos presenta valores inferiores, lo cual resulta favorable para edificios climatizados mediante aire acondicionado.
- La transmitancia térmica promedio del acristalamiento (UVG) es -4,31 $W/m^2 K$, mientras que el coeficiente de ganancia solar (SHGC) y la transmitancia luminosa visible (VLT) promedian 0,54 y 0,66, respectivamente. En otras palabras, el vidrio permite pasar, en promedio,

alrededor del 54 % de la radiación solar incidente y el 66 % de la luz visible, bloqueando el resto. Las medianas correspondientes son 5,05 $W/m^2 K$, 0,50 y 0,73, lo que implica que la mitad de los proyectos posee valores de UVG iguales o inferiores a 5,05 $W/m^2 K$ y SHGC iguales o inferiores a 0,50, mientras que la VLT es igual o superior a 0,73.

- El porcentaje de ahorro energético (PAH) presenta un valor medio del 21 %, mientras que la mediana alcanza el 22 %. Por lo tanto, la mitad de los proyectos registra un ahorro igual o inferior al 22 %.
- La variable con mayor variabilidad relativa es ERN (*Energía Renovable Neta*, kWh/m^2 año), que mide la fracción de la demanda energética anual cubierta por renovables instaladas *in situ*. Su coeficiente de variación (CV) alcanza el 54,3 %, lo que evidencia una notable dispersión respecto de la media.

A continuación se describe la distribución de las variables cualitativas TIP, CIU, ZCL, CER.

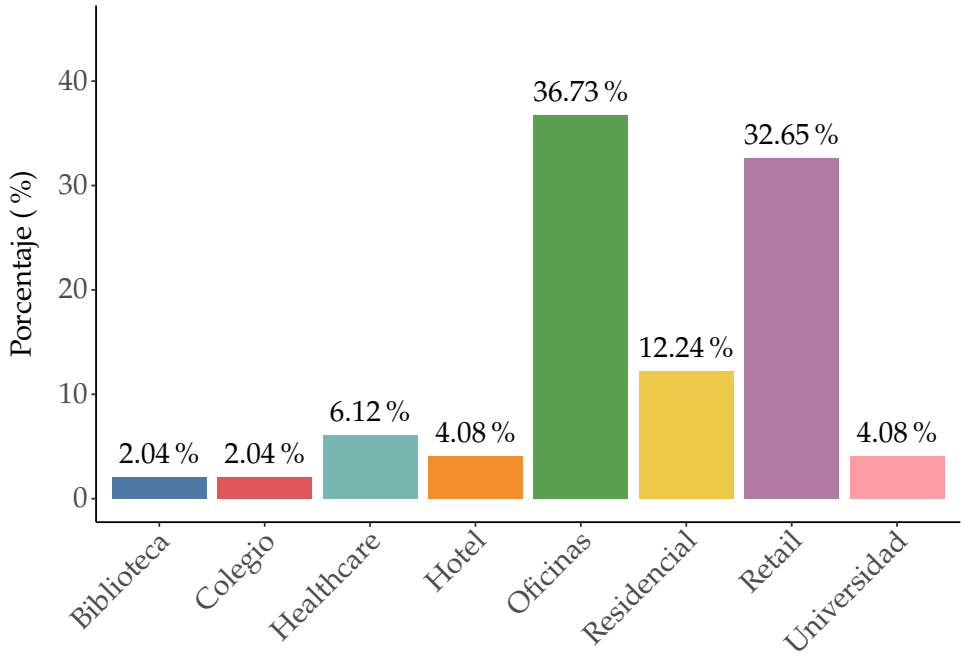

Figura 2: Distribución porcentual por tipo de proyecto (TIP)

- El 69 % de los proyectos son de oficinas y retail. Esto se debe a que este tipo de proyectos normalmente cuenta con un mayor capital que

las otras tipologías de proyectos. Al tener mayor capital, los proyectos constructivos pueden adquirir más servicios, tales como optar por una certificación sostenible y así tener un modelo energético (ver la Figura 2).

- El 51 % de los proyectos se centran en Bogotá, puesto que es la capital y presenta un crecimiento mayor de proyectos constructivos, seguido de Medellín con un 12,24 %, que es otra ciudad principal de Colombia (ver la Figura 3).

- La clasificación de zonas climáticas usada en los modelos energéticos cubre el rango de 1A a 4C. Las categorías 3A, 3C y 4C, habitualmente asociadas a Bogotá y a climas fríos, concentran el 63,27 % de los proyectos analizados. En cambio, la categoría 2A se aplica a climas cálido−templados, como el de Medellín, según la normativa *ASHRAE* (ver la Figura 4).

- Es importante tener en cuenta que la clasificación de zona climática asignada a una ciudad puede variar, pues depende del criterio de quien elabora el modelo y del año de referencia de la norma *ASHRAE* que se utilice.

- Uno de los 49 proyectos analizados se ubica en Quito (Ecuador). Según la norma *ASHRAE* 90.1 (2016), esta ciudad pertenece a la zona climática 3A, la misma que se asigna a Bogotá; por ello se incluye en las categorías predominantes del conjunto colombiano. En consecuencia, dicho proyecto aporta evidencia adicional y contribuye a compensar la limitada cantidad de casos disponibles en este estudio (ver la Figura 4).

- El 10,12 % de los proyectos figura con certificación *N/A*; esto significa que no obtuvo ninguna certificación porque no alcanzó el requisito mínimo de un ahorro energético del 5 % con respecto a la línea base establecida por *ASHRAE* (véase la Figura 5).

- La certificación $LEED\ v4 - BD + C :\ NC$, dirigida a proyectos de nueva construcción, es la más frecuente, con un aproximado de 26,53 % de los casos (ver la Figura 5).

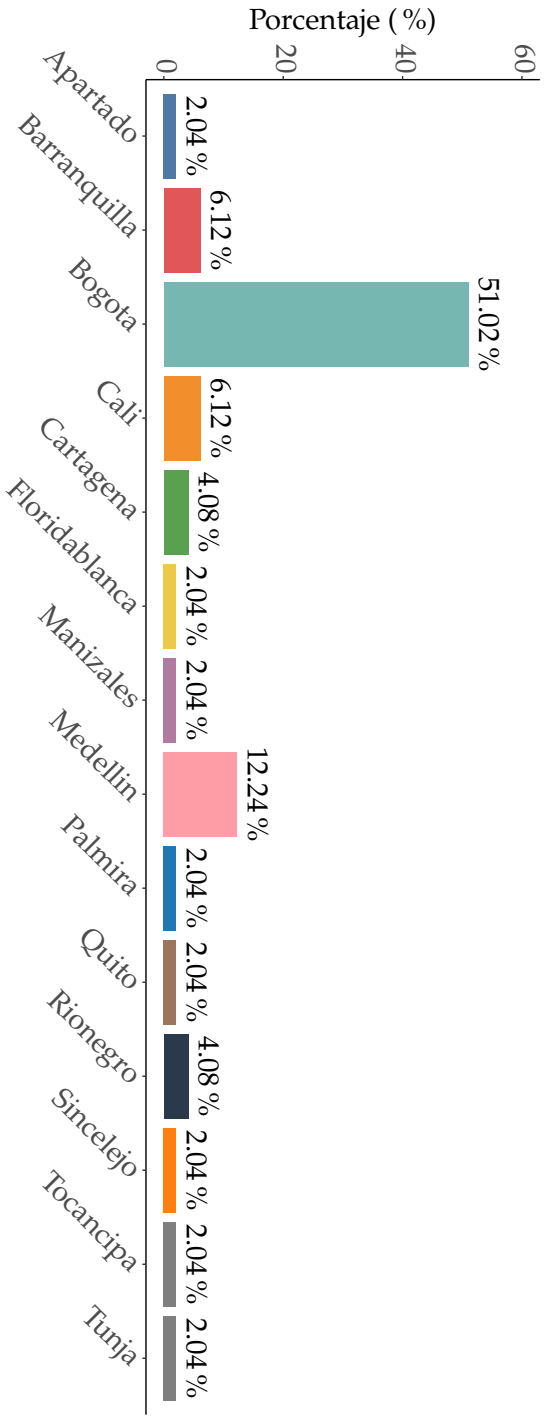

Figura 3: Distribución porcentual por ciudad (CIU)

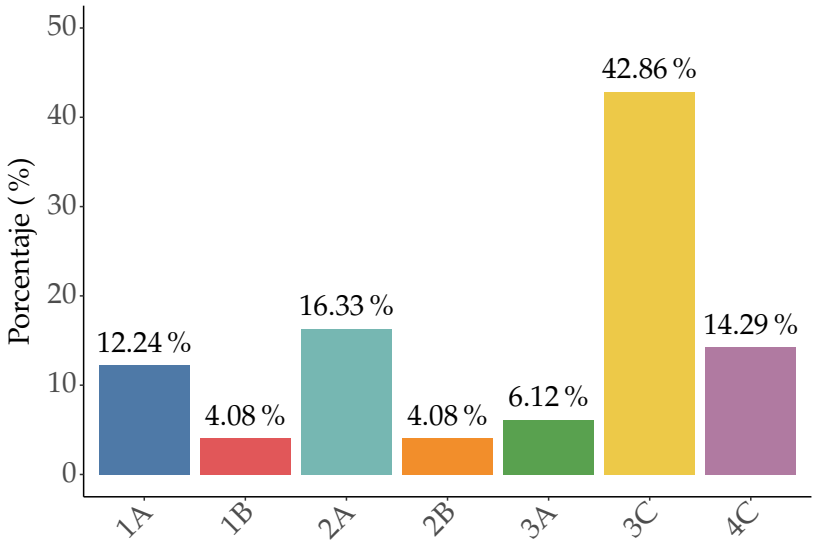

Figura 4: Distribución porcentual por zona climática (ZCL)

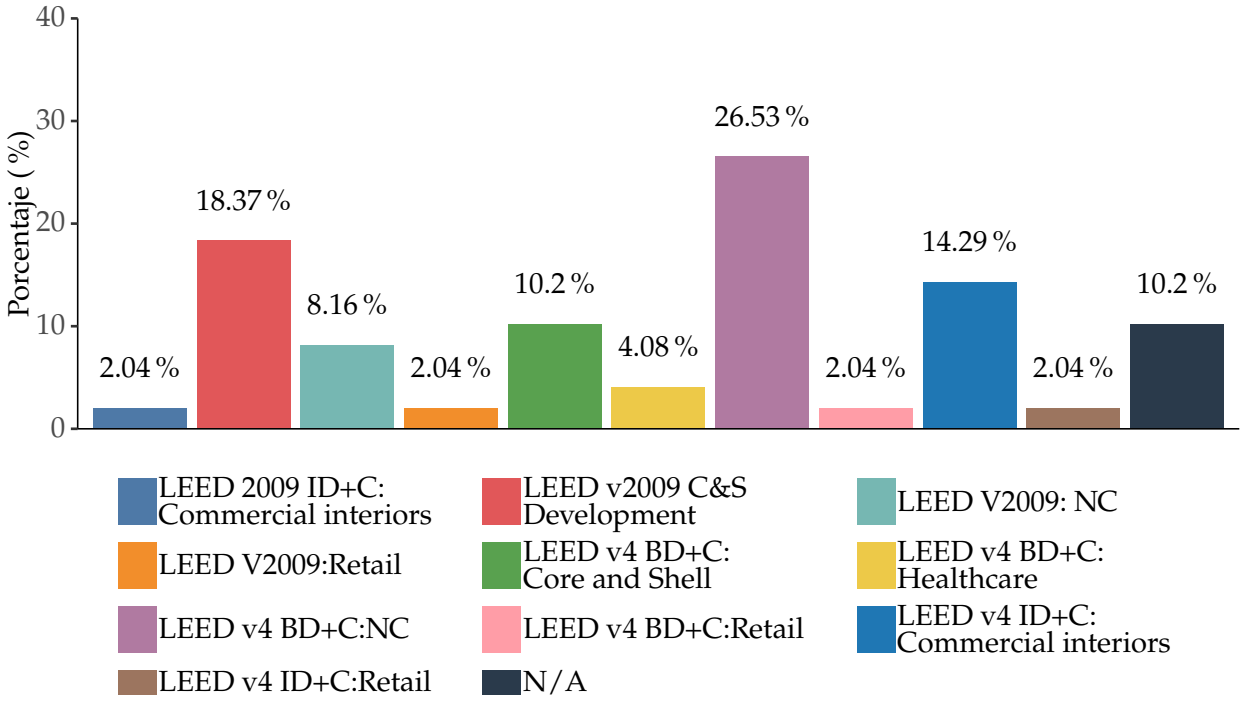

Figura 5: Distribución porcentual por tipo de certificación (CER)

A continuación se describe la distribución de las variables cuantitativas ART y AAC.

Tabla 3: Resumen estadístico de la variable ART

Estadística	Valor
Mínimo	79,10
1er Cuartil	2983,00
Mediana	7441,35
Media	27734,17
3er Cuartil	31264,00
Máximo	236009,70
Coeficiente de Variación (%)	178,16

El Área de Referencia Térmica (ART) presenta una dispersión muy alta respecto a su media: su coeficiente de variación alcanza el 178,16 %. Los valores oscilan entre 79,1 m^2 y 236 000 m^2. El tercer cuartil se sitúa en 31 264 m^2, de modo que el 75 % de los proyectos posee un ART igual o inferior a dicha superficie (véase la Tabla 10).

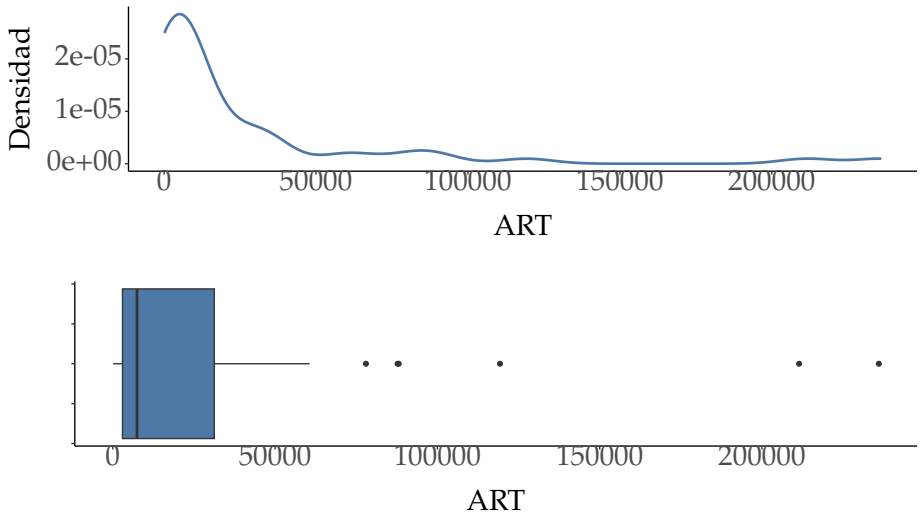

Figura 6: Gráficos estadísticos variable ART

Entre los proyectos analizados, seis presentan valores atípicos de Área de Referencia Térmica (ART), comprendidos entre 77 996 m^2 y 236 000 m^2, como se muestra en el diagrama de caja de la Figura 6. Estos casos corresponden a un edificio de tipo *Healthcare* y cinco de tipo *Retail*. En cuanto a

su ubicación, los proyectos atípicos se distribuyen del siguiente modo: dos en Bogotá y uno en Manizales, Sincelejo, Barranquilla y Medellín, respectivamente.

Adicionalmente, la distribución del ART exhibe un sesgo positivo; la curva de densidad revela una marcada concentración de proyectos por debajo de 50 000 m^2, mientras que son escasos los que superan esa superficie (véase la Figura 6).

Tabla 4: Resumen estadístico de la variable ART sin atípicos

Estadística	Valor
Mínimo	79,10
1er Cuartil	2557,25
Mediana	6691,49
Media	12159,12
3er Cuartil	12042,28
Máximo	60629,90
Coeficiente de Variación (%)	123,70

Al excluir los seis valores atípicos de ART, el coeficiente de variación se reduce al 124,0 % (véase la Tabla 12), lo que supone una disminución notable. No obstante, el área de los proyectos sigue mostrando una variabilidad considerable respecto de su media.

Tabla 5: Resumen estadístico de la variable AAC

Estadística	Valor
Mínimo	0,00
1er Cuartil	2027,86
Mediana	5621,50
Media	15111,02
3er Cuartil	18019,93
Máximo	117705,60
Coeficiente de Variación (%)	157,54

La superficie climatizada (*AAC*) exhibe un rango muy amplio, desde 0 m^2 hasta 117 706 m^2. Un valor de 0 m^2 indica que el proyecto recurre únicamente a ventilación natural para garantizar el confort térmico. El tercer cuartil se ubica en 18 722 m^2, por lo que el 75 % de los edificios presenta un

AAC igual o inferior a esa superficie. La variabilidad relativa es elevada: el coeficiente de variación alcanza el 157 %. Además, la mediana (5 622 m^2) es sensiblemente menor que la media (15 111 m^2), lo que confirma un sesgo positivo en la distribución (ver la Tabla 5)

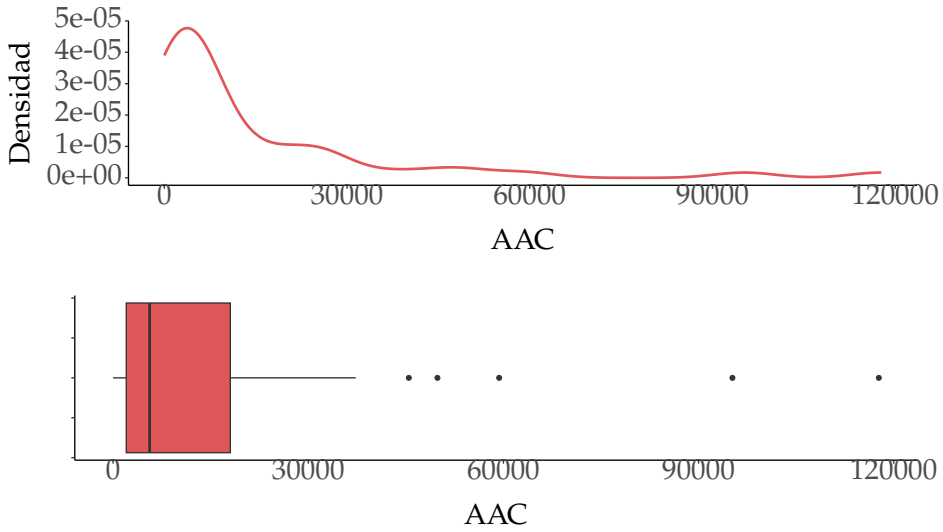

Figura 7: Gráficos estadísticos variable AAC)

Cinco proyectos presentan valores atípicos de superficie climatizada (*AAC*), comprendidos entre 49 459 m^2 y 117 706 m^2, como se aprecia en el diagrama de caja de la Figura 7. Todos corresponden al sector *retail* y se distribuyen del siguiente modo: uno en Manizales, uno en Sincelejo, dos en Bogotá y uno en Medellín. La curva de densidad revela un sesgo positivo, con una marcada concentración de proyectos por debajo de 30 000 m^2.

Tabla 6: Resumen estadístico de la variable AAC sin atípicos

Estadística	Valor
Mínimo	0,00
1er Cuartil	1468,00
Mediana	5508,60
Media	8319,98
3er Cuartil	9374,66
Máximo	37300,00
Coeficiente de Variación (%)	111,68

Al excluir los cinco valores atípicos de AAC, el coeficiente de variación

se reduce al 112,0 % (véase la Tabla 6), lo que supone una disminución importante. Sin embargo, la superficie climatizada de los proyectos sigue mostrando una variabilidad considerable respecto de su media.

2. *Análisis Bivariado*

Se efectuaron análisis bivariados para identificar relaciones redundantes entre variables que pudieran sesgar los resultados de los modelos y métodos empleados en este estudio.

Tabla 7: Distribución de Tipología por Ciudad

	Biblioteca	Colegio	Healthcare	Hotel	Oficinas	Residencial	Retail	Universidad	Totales
Apartado	0	0	0	0	0	1	0	0	1
Barranquilla	1	0	0	0	1	0	1	0	3
Bogotá	0	1	2	0	14	3	4	1	25
Cali	0	0	0	1	0	0	1	1	3
Cartagena	0	0	0	1	0	1	0	0	2
Floridablanca	0	0	0	0	0	0	1	0	1
Manizales	0	0	0	0	0	0	1	0	1
Medellín	0	0	1	0	1	1	3	0	6
Palmira	0	0	0	0	0	0	1	0	1
Quito	0	0	0	0	1	0	0	0	1
Rionegro	0	0	0	0	1	0	1	0	2
Sincelejo	0	0	0	0	0	0	1	0	1
Tocancipá	0	0	0	0	0	0	1	0	1
Tunja	0	0	0	0	0	0	1	0	1
Totales	**1**	**1**	**3**	**2**	**18**	**6**	**16**	**2**	**49**

Tabla 8: Distribución de Zona Climática por Ciudad

	1A	1B	2A	2B	3A	3C	4C	Totales
Apartado	1	0	0	0	0	0	0	1
Barranquilla	1	2	0	0	0	0	0	3
Bogotá	0	0	1	0	1	18	5	25
Cali	1	0	0	2	0	0	0	3
Cartagena	2	0	0	0	0	0	0	2
Floridablanca	0	0	1	0	0	0	0	1
Manizales	0	0	0	0	0	1	0	1
Medellín	0	0	5	0	1	0	0	6
Palmira	0	0	1	0	0	0	0	1
Quito	0	0	0	0	1	0	0	1
Rionegro	0	0	0	0	0	2	0	2
Sincelejo	1	0	0	0	0	0	0	1
Tocancipá	0	0	0	0	0	0	1	1
Tunja	0	0	0	0	0	0	1	1
Totales	**6**	**2**	**8**	**2**	**3**	**21**	**7**	**49**

Tabla 9: Distribución de Certificación por Ciudad

Ciudad	LEED_2009 ID+C: Commercial interiors	LEED_v2009 C&S Development	LEED_v2009: NC	LEED_v2009: Retail	LEED_v4_BD+C: Core and Shell	LEED_v4_BD+C: Healthcare	LEED_v4_BD+C: NC	LEED_v4_BD+C: Retail	LEED_v4_ID+C: Commercial interiors	LEED_v4_ID+C: Retail	N/A	Totales
Apartado	0	0	0	0	0	0	0	0	0	0	1	1
Barranquilla	0	1	1	0	1	0	0	0	0	0	0	3
Bogotá	1	4	1	1	2	2	6	1	5	1	1	25
Cali	0	0	0	0	1	0	1	0	0	0	1	3
Cartagena	0	0	0	0	0	0	1	0	0	0	1	2
Floridablanca	0	0	1	0	0	0	0	0	0	0	0	1
Manizales	0	1	0	0	0	0	0	0	0	0	0	1
Medellín	0	2	0	0	1	0	2	0	1	0	0	6
Palmira	0	0	0	0	0	0	1	0	0	0	0	1
Quito	0	0	0	0	0	0	0	0	1	0	0	1
Rionegro	0	0	0	0	0	0	1	0	0	0	1	2
Sincelejo	0	1	0	0	0	0	0	0	0	0	0	1
Tocancipá	0	0	0	0	0	0	1	0	0	0	0	1
Tunja	0	0	1	0	0	0	0	0	0	0	0	1
Totales	**1**	**9**	**4**	**1**	**5**	**2**	**13**	**1**	**7**	**1**	**5**	**49**

Luego de realizar el análisis de la base de datos en función de la variable CIU, se puede determinar lo siguiente:

a) **TIP vs CIU (ver la Tabla 7)**

- En la ciudad de Bogotá, las Oficinas son el tipo de edificación más certificada con un 56 %. Las categorías Retail y Residencial le siguen con 16 % y 12 %, respectivamente.
- En Medellín, el tipo de edificación con mayor porcentaje de certificación es Retail con 50 %. El porcentaje restante se reparte de manera equitativa en Healthcare, Oficinas y Residencial.
- En Barranquilla y Cali, las certificaciones están distribuidas equitativamente en un 33,33 % entre las categorías Biblioteca, Oficinas y Retail.
- De forma general, la categoría Retail está presente al menos una vez en el 70 % de las ciudades y es destacada en Bogotá y Medellín, con 4 y 3 apariciones respectivamente.

b) **ZCL vs CIU (ver la Tabla 8)**

- En Bogotá, el 85,7 % de las edificaciones tienen clasificación 3C, seguido por 4C con 71,4 %.
- Medellín concentra el 62,5 % de las edificaciones en zonas climáticas de clase 2A.
- La asignación de zona climática depende del criterio del certificador, lo que podría introducir sesgos; sería importante profundizar en esta asignación.

c) **CER vs CIU (ver la Tabla 9)**

- En Bogotá, la certificación más frecuente es $LEED-v4-BD+C : NC$ con 24 %, seguida por $LEED-v4-ID+C : CommercialInteriors$ (20 %) y $LEED - v2009 - C\&S - Development$ (16 %).
- En Medellín, tanto $LEED - v2009 - C\&S - Development$ como $LEED - v4 - BD + C : NC$ representan un 33,33 % cada una. El restante 33,33 % se reparte entre $LEED - v4 - ID + C : Commercial - Interiors$ y $LEED - v4 - BD + C : Core - and - Shell$.
- En Barranquilla, las certificaciones $LEED-V2009 : NC$, $LEED-v2009 - C\&S - Development$ y $LEED - v4 - BD + C : Core$ and Shell se distribuyen equitativamente (33,33 % cada una).

- Los proyectos sin certificación se encuentran en Bogotá, Cali, Cartagena y Rionegro.

A continuación se analizan en conjunto los variables ART y AAC.

Las variables Área Total (ART) y Área Acondicionada (AAC) presentan comportamientos estadísticos y gráficos similares, lo cual es esperable, ya que un mayor ART suele implicar un mayor requerimiento de área acondicionada (ver las Figuras 8, 9 y la Tabla 10).

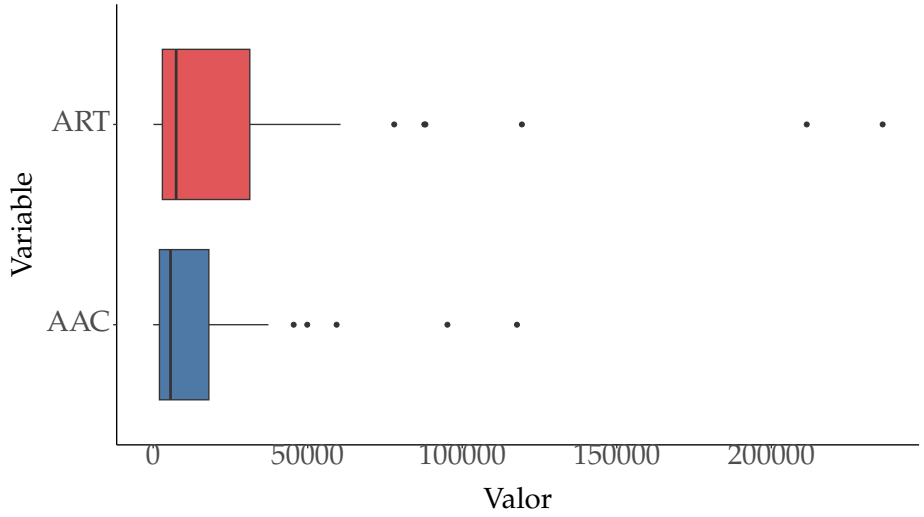

Figura 8: Comparación de ART y AAC

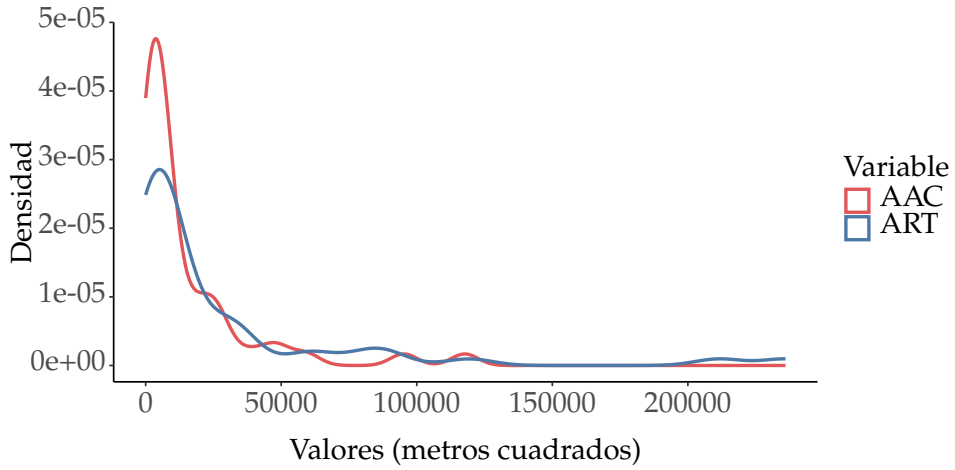

Figura 9: comparativo de Densidad variables ART y AAC

Tabla 10: Resumen estadístico de la variable ART

Estadística	Valor
Mínimo	79,10
1er Cuartil	2983,00
Mediana	7441,35
Media	27734,17
3er Cuartil	31264,00
Máximo	236009,70
Coeficiente de Variación (%)	178,16

Tabla 11: Resumen estadístico de la variable AAC

Estadística	Valor
Mínimo	0,00
1er Cuartil	2027,86
Mediana	5621,50
Media	15111,02
3er Cuartil	18019,93
Máximo	117705,60
Coeficiente de Variación (%)	157,54

a) Los diagramas de caja (ver la Figura 8) evidencian la presencia de proyectos atípicos en ambas variables:

- Para ART, se identifican 6 proyectos atípicos (1 de tipo *healthcare* y 5 *retail*), ubicados en Bogotá (2), Manizales (1), Sincelejo (1), Barranquilla (1) y Medellín (1), con valores entre 77996 m^2 y 236000 m^2.
- En AAC, se encuentran 5 proyectos atípicos (todos *retail*) en Manizales, Sincelejo, Bogotá (2) y Medellín, con áreas entre 49459 m^2 y 117706 m^2.

b) Las curvas de densidad (ver la Figura 9) muestran que:

- En ART, la mayoría de los proyectos se concentran por debajo de 100000 m^2, mientras que los valores superiores corresponden a los proyectos atípicos.
- En AAC, la concentración de proyectos se da en áreas menores a 50000 m^2, manteniendo un patrón similar al de ART, aunque con una escala más reducida.

c) En el caso del ART, los proyectos analizados presentan un rango entre 79,1 m^2 y 236000 m^2, mientras que el AAC varía desde 0 m^2 (proyectos que utilizan ventilación natural) hasta 117706 m^2. El 75 % de los proyectos tienen un ART menor o igual a 31264 m^2 y un AAC menor o igual a 18722 m^2 (ver la Tabla 10).

d) Ambas variables presentan una alta variabilidad respecto a su media:

- En ART, el coeficiente de variación es 178,16 %, con una media de 27734,2 m^2 y una mediana considerablemente menor (7441,4 m^2), indicando una distribución altamente sesgada a la derecha.

- En AAC, el coeficiente de variación es 157 %, con una media de 15111 m² y una mediana de 5622 m², también con un sesgo positivo evidente.

e) En conclusión, aunque ART y AAC son variables distintas, sus comportamientos estadísticos y gráficos están alineados, reflejando una relación estructural entre el tamaño total de los proyectos y el área acondicionada que requieren. Esta similitud debe ser tenida en cuenta en análisis posteriores, especialmente al modelar relaciones entre variables o detectar redundancias.

Tras eliminar los valores atípicos de las variables AAC (5 datos) y ART (6 datos), se observa:

Tabla 12: Resumen estadístico de la variable ART sin atípicos

Estadística	Valor
Mínimo	79,10
1er Cuartil	2557,25
Mediana	6691,49
Media	12159,12
3er Cuartil	12042,28
Máximo	60629,90
Coeficiente de Variación (%)	123,70

Tabla 13: Resumen estadístico de la variable AAC sin atípicos

Estadística	Valor
Mínimo	0,00
1er Cuartil	1468,00
Mediana	5508,60
Media	8319,98
3er Cuartil	9374,66
Máximo	37300,00
Coeficiente de Variación (%)	111,68

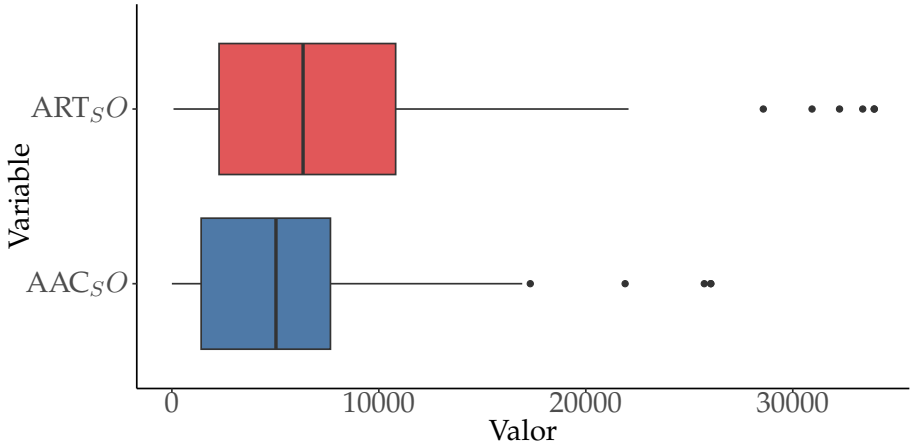

Figura 10: Comparativo box-plot variables ART y AAC sin atípicos

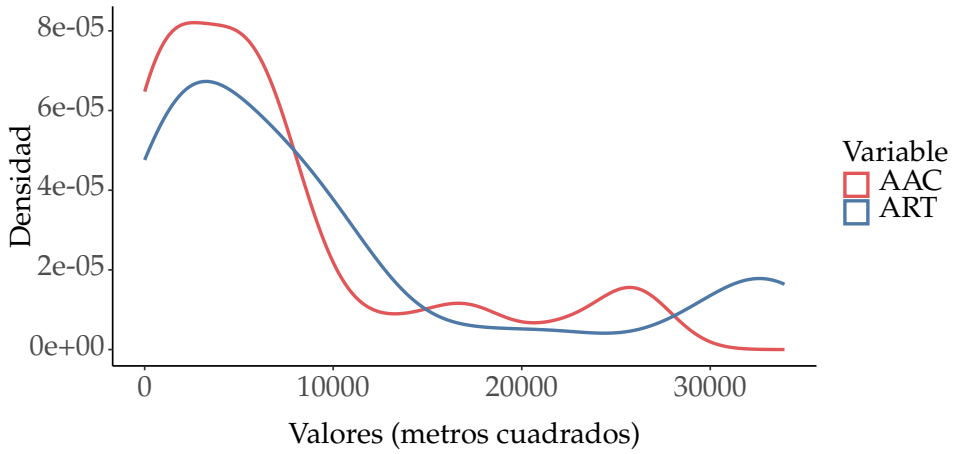

Figura 11: comparativo de densidad variables ART y AAC sin atipicos

a) Una reducción significativa en su coeficiente de variación: del ART pasa a 124 % y del AAC a 112 % (ver la Tabla 12). Aunque estas reducciones son considerables, ambas variables continúan presentando una alta dispersión respecto a su media, lo cual indica una notable variabilidad entre los proyectos analizados.

b) Los gráficos de cajas comparativos evidencian al menos cuatro valores atípicos en ambas variables (ver la Figura 10). Para el AAC, los valores oscilan entre 0 y 26040 m^2, lo que puede asociarse a edificaciones que utilizan ventilación natural como modo principal de acondicionamiento. A pesar de eliminar los atípicos, la dispersión sigue siendo considerable, manteniéndose los coeficientes de variación por encima del 100 %.

c) El análisis conjunto de las densidades (ver la Figura 11) sin valores atípicos permite visualizar mejor la distribución de ambas variables. Se evidencia que tanto el AAC como el ART se mantienen por debajo de los 35000 m^2. No obstante, el AAC presenta picos de frecuencia entre los 25000 m^2 y 30000 m^2, mientras que el ART tiene mayor concentración entre los 25000 m^2 y 35000 m^2. En ambos casos, se destaca una mayor concentración de proyectos con áreas menores a 20000 m^2.

A continuación se muestra la distribución de proyectos distribuidos según su zona climática y ciudad:

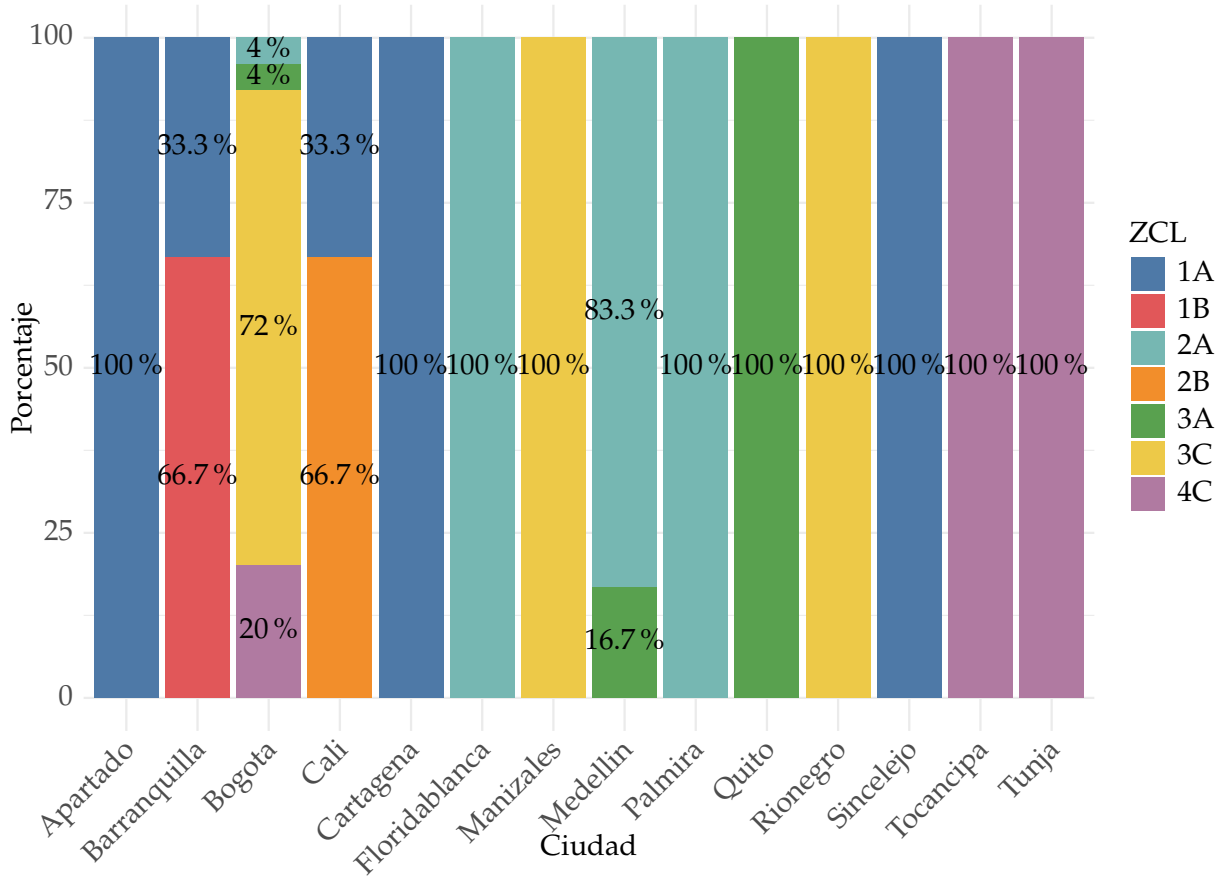

Figura 12: Distribución Zona Climática por Ciudad

a) Bogotá figura con las zonas climáticas 2A, 3A, 3C y 4C. A partir de la edición 2016 del estándar *ASHRAE 90.1* se oficializó la clasificación de la ciudad como 3A (tibia-húmeda); por ello, se presume que los proyectos asignados a otras zonas fueron elaborados antes de esa fecha, lo que explica la dispersión observada. En la base de datos, la zona más frecuente es la 3C (72 %), mientras que las menos representadas son 2A y 3A, cada una con apenas el 4 % (ver la Figura 12).

b) Adicionalmente. los climas de Manizales y Rionegro son muy similares al de Bogotá; en el gráfico de barras ambos aparecen en color naranja y se agrupan en la zona climática 3C. Por el contrario, las zonas 1A (muy caliente húmeda) y 1B (muy caliente seca) son propias de regiones costeras, por lo que Cali no debería figurar con esa clasificación. Asimismo, la zona climática 2A (caliente húmeda) corresponde a Medellín, Cali y Palmira, de modo que Bogotá no debería incluirse

en esa categoría.

Seguidamente se describen en conjunto algunas variables cualitativas con algunas variables cuantitativas.

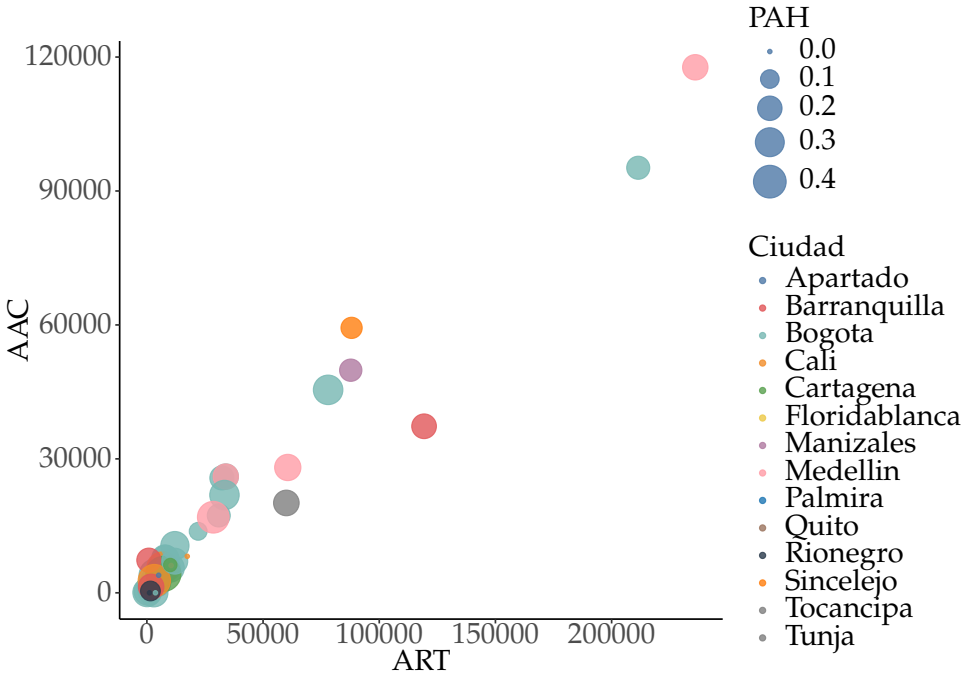

Figura 13: Dispersión ART, AAC, PAH

a) La relación entre las variables ART y AAC, con el tamaño de los círculos proporcional al porcentaje de ahorro (PAH). Se observa una tendencia clara: cuanto menores son los valores de ART y AAC, mayor es el ahorro, lo que se refleja en los círculos de mayor diámetro concentrados en la esquina inferior izquierda del gráfico. En cambio, los proyectos con ART y AAC más elevados, ubicados principalmente en Bogotá y Medellín, presentan porcentajes de ahorro inferiores al 20 % (ver la Figura 13).

b) En la Figura 14 se presenta la relación de las variables AFO, ORN teniendo en cuenta el valor de CPR. Los proyectos orientados a menos de $150°$ con respecto al norte y con un área de fachada inferior a $2\,500\,m^2$ presentan un menor consumo energético. Asimismo, los edificios cuya orientación se sitúa entre $200°$ y $350°$ y cuya fachada no supera los $1\,250\,m^2$ también registran un consumo reducido. La base de datos no contiene proyectos con orientaciones comprendidas

entre 150° y 200°. Finalmente, las edificaciones orientadas entre 250° y 360° y con más de 2 500 m^2 de superficie de fachada oriental muestran valores más altos de *CPR*.

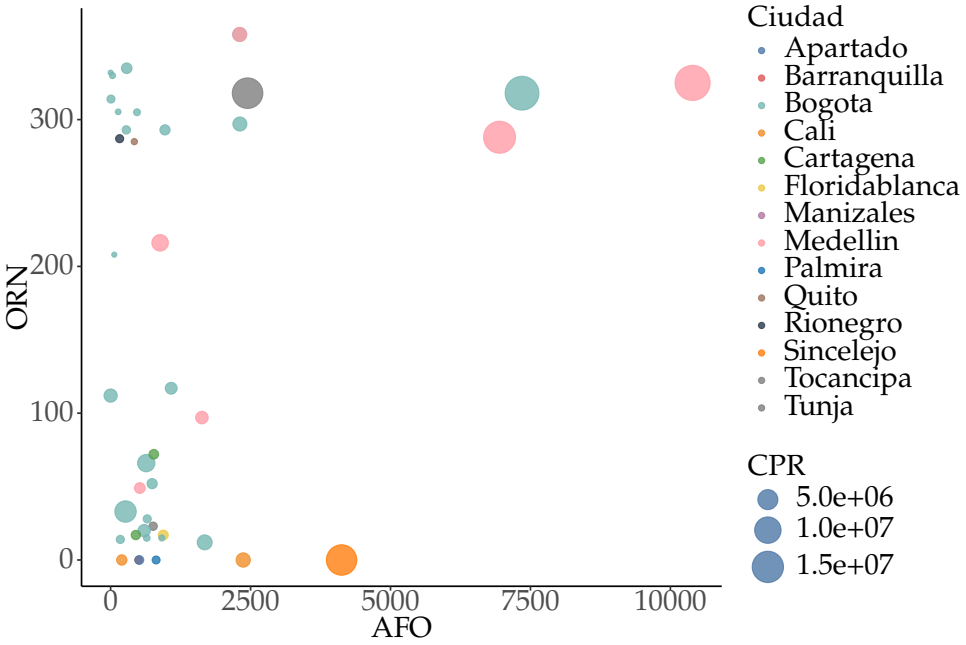

Figura 14: Dispersión AFO, ORN, CPR

c) En la Figura 15 se presenta la relación de las variables WWR, AAC teniendo en cuenta el CPR. Cuando el AAC se mantiene por debajo de 15 000 m^2, el CPR permanece bajo con independencia del WWR. El AAC está estrechamente ligado al CPR y, cuanto mayor es el WWR, mayor tiende a ser este indicador. La tipología con los valores más altos de CPR es la de retail, probablemente debido a su gran tamaño y a la elevada densidad de ocupación. Por su parte, los edificios de oficinas exhiben los WWR más altos, aunque ninguno supera los 30 000 m^2 de AAC.

d) A continuación, en la Figura 16 se presenta la relación de las variables ACB, PAH teniendo en cuenta el CPR. Los proyectos con valores más altos de ACB también registran los CPR más elevados, independientemente de la ciudad, lo que confirma una relación directa entre ambas variables.

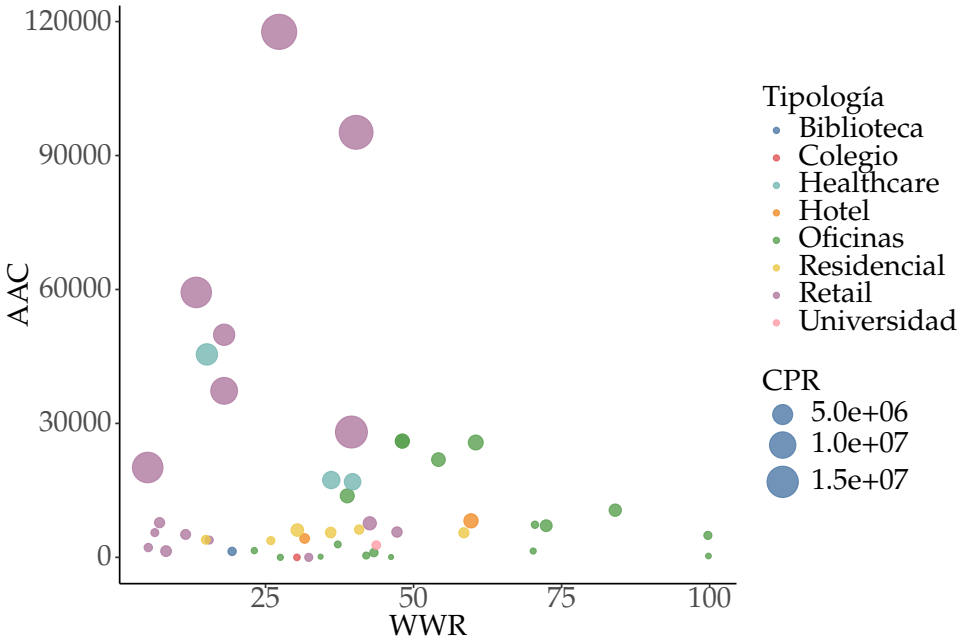

Figura 15: Dispersión WWR, AAC, CPR

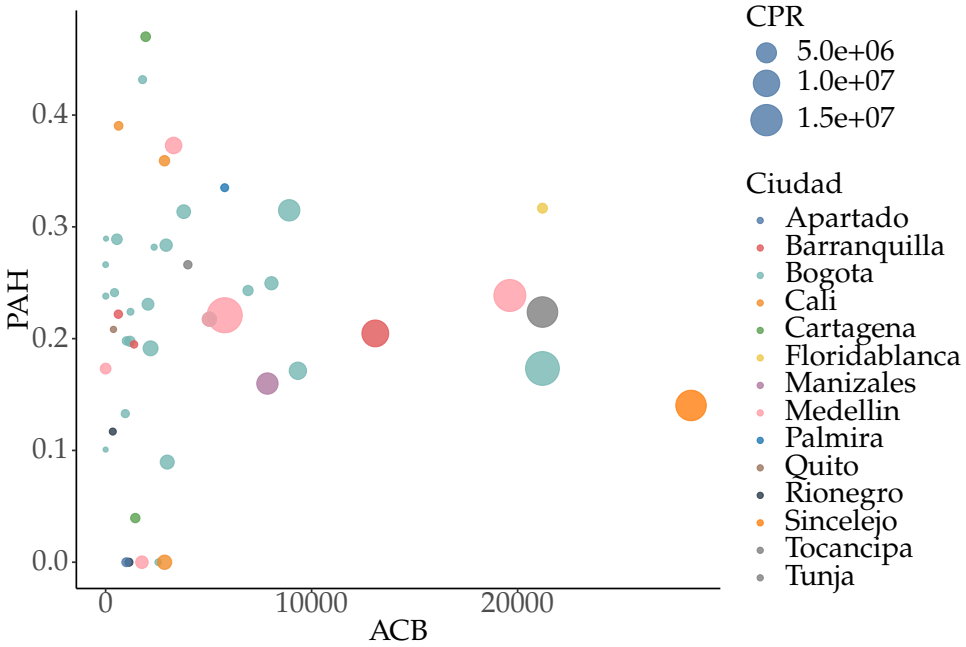

Figura 16: Dispersión ACB, PAH, CPR

e) De manera general, los casos cuyo PAH supera el 25 % presentan un ACB inferior a 1 000 m^2; la única excepción corresponde a un proyecto en Floridablanca, con un ACB cercano a 20 000 m^2 y un PAH de alrededor del 33 %. Este comportamiento atípico podría explicarse por la integración de paneles fotovoltaicos o por estrategias bioclimáticas que reducen la ganancia de radiación solar. En contraste, ningún proyecto con ACB superior a 1 000 m^2 supera el 25 % de PAH.

3. *Datos Faltantes*

La Figura 17 muestra el diagrama de datos ausentes, que resume de forma gráfica la cantidad y distribución de valores faltantes en la base de datos. En la parte superior se listan las variables y, debajo de cada nombre, se indica el número de ausencias en esa columna. A la izquierda se presenta el número de proyectos que comparten un mismo patrón de ausencia; por ejemplo, la primera fila agrupa los 33 proyectos completos y la segunda fila los 3 proyectos que carecen únicamente de la variable de orientación norte. Los números ubicados a la derecha representan el total de variables faltantes por fila. Por último, la cifra de la esquina inferior derecha corresponde a la suma global de valores faltantes, que en este caso asciende a 47.

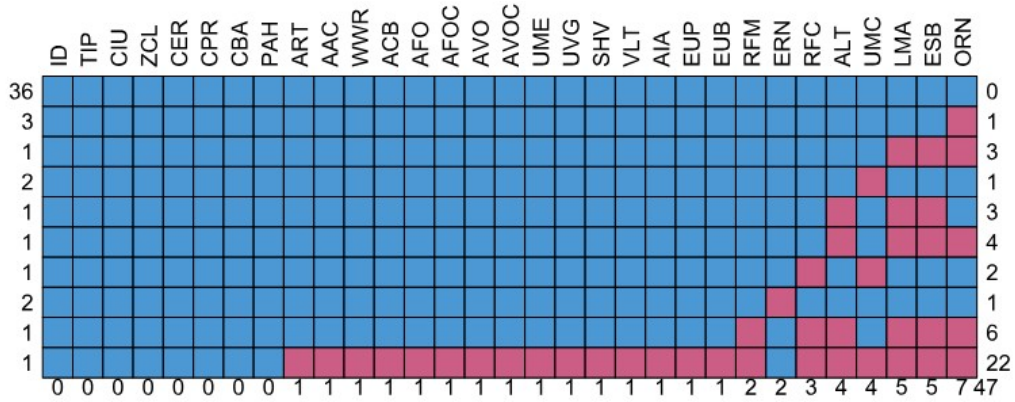

Figura 17: Identificación de faltantes por variable

Respecto a la Figura 17:

- La base contiene 47 valores faltantes.
- Treinta y seis proyectos están completos, sin ausencias.
- La variable ORN presenta 7 valores faltantes; le siguen ESB y LMA con 5 cada una.

- Un proyecto carece de 22 variables, probablemente por tratarse de un registro antiguo con prácticas de recolección distintas.
- Otro proyecto muestra 6 valores faltantes por razones similares.
- Tres proyectos tienen una sola ausencia correspondiente a ORN.
- Un proyecto presenta 3 ausencias: LMA, ESB y ORN.
- Dos proyectos carecen de UMC, quizá porque se localizan en pisos intermedios o la información no estaba disponible.
- Uno carece de ALT, LMA y ESB.
- Otro carece de ALT, LMA, ESB y ORN.
- Finalmente, un proyecto carece de ALT y UMC.

No se aprecia un patrón sistemático en la distribución de los datos faltantes. Los datos faltan *completamente al azar (MCAR)*. Esto implica que las posibles causas de los datos faltantes no están relacionadas con los valores de los datos. Por lo tanto, podemos ignorar muchas de las complejidades que surgen debido a la falta de datos.

El proceso de imputación en las variables cuantitativas con valores faltantes se realizó utilizando la *mediana* debido al sesgo pronunciado de las variables cuantitativas. Después de este procedimiento ninguna variable tiene valores faltantes. Los resultados de las estadísticas antes y después de la imputación son muy similares, lo que refleja una buena decisión en el proceso de imputación.

4. *Análisis de Correlaciones*

   Se empleó el coeficiente de correlación de Kendall, en lugar del habitual de Pearson, para detectar posibles asociaciones entre las variables cuantitativas. Esta elección se justifica porque dichas variables presentan distribuciones sesgadas, por lo que un método no paramétrico como el de Kendall se ajusta mejor a la naturaleza de los datos y al objetivo del estudio. A continuación se muestran los resultados obtenidos con este enfoque (ver la Figura 18):

   - Existe una correlación positiva fuerte entre la ALT y la ESB. Esto se debe a que la ESB es una relación entre el LMA y la ALT.
   - El ART y el AAC tienen una correlación positiva fuerte con el AFO y el AFOC. Esto se debe a que edificios más grandes tendrán más área orientada hacia el oriente y hacia el occidente.
   - El ART y el AAC tienen una correlación positiva fuerte con el CPR y el CBA. Esto se debe a que edificios más grandes tendrán mayor CPR.

- El AFOC tiene una correlación positiva fuerte con AFO, CBA y CPR. Esto indica que entre más AFOC se tenga, habrá más AFO y, por lo tanto, mayor CPR.
- El AVOC tiene una correlación positiva fuerte con AVO, CBA y CPR. Esto indica que entre más AVOC se tenga, habrá más AVO y, por lo tanto, mayor CPR.

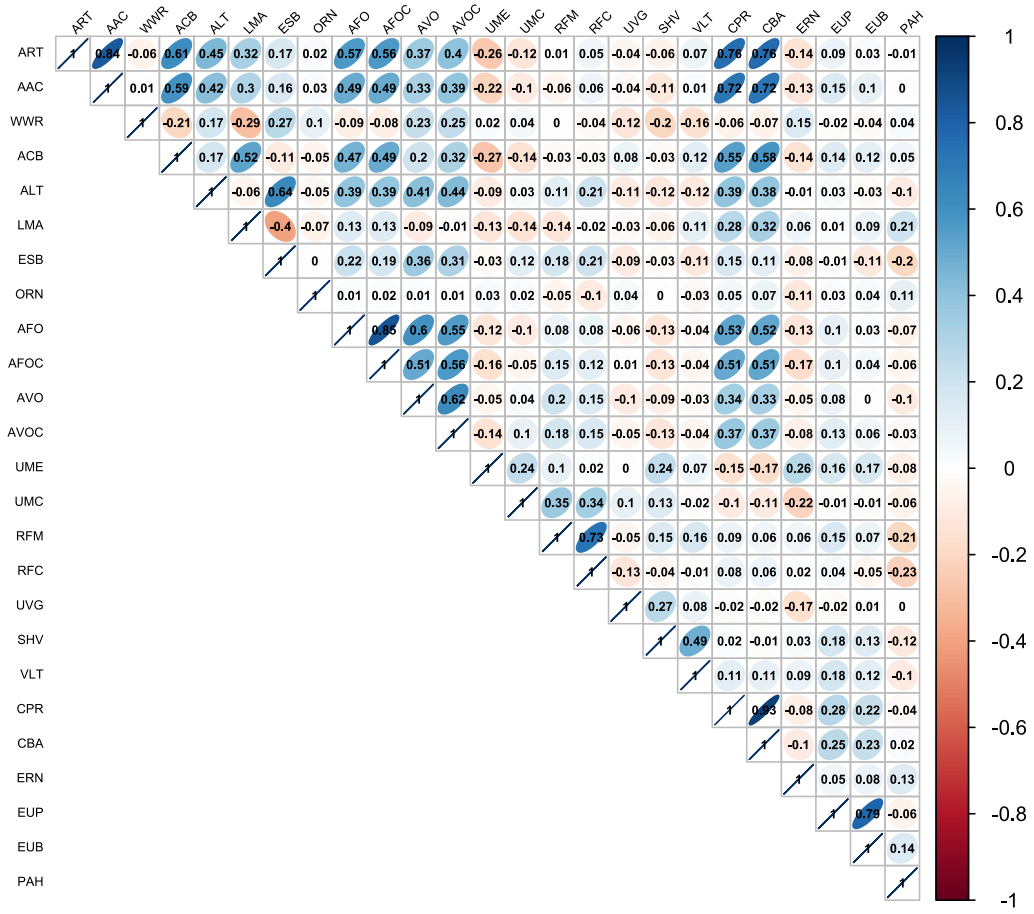

Figura 18: Análisis de correlación de Kendall

Posteriormente a este análisis se eliminaron aquellas variables que pueden resultar redundantes y generar ruido en el modelo estadístico. Estas variables son: ART, ESB, AFOC, AVOC, CPR, CBA, ERN, EUP y EUB (ver la Figura 19).

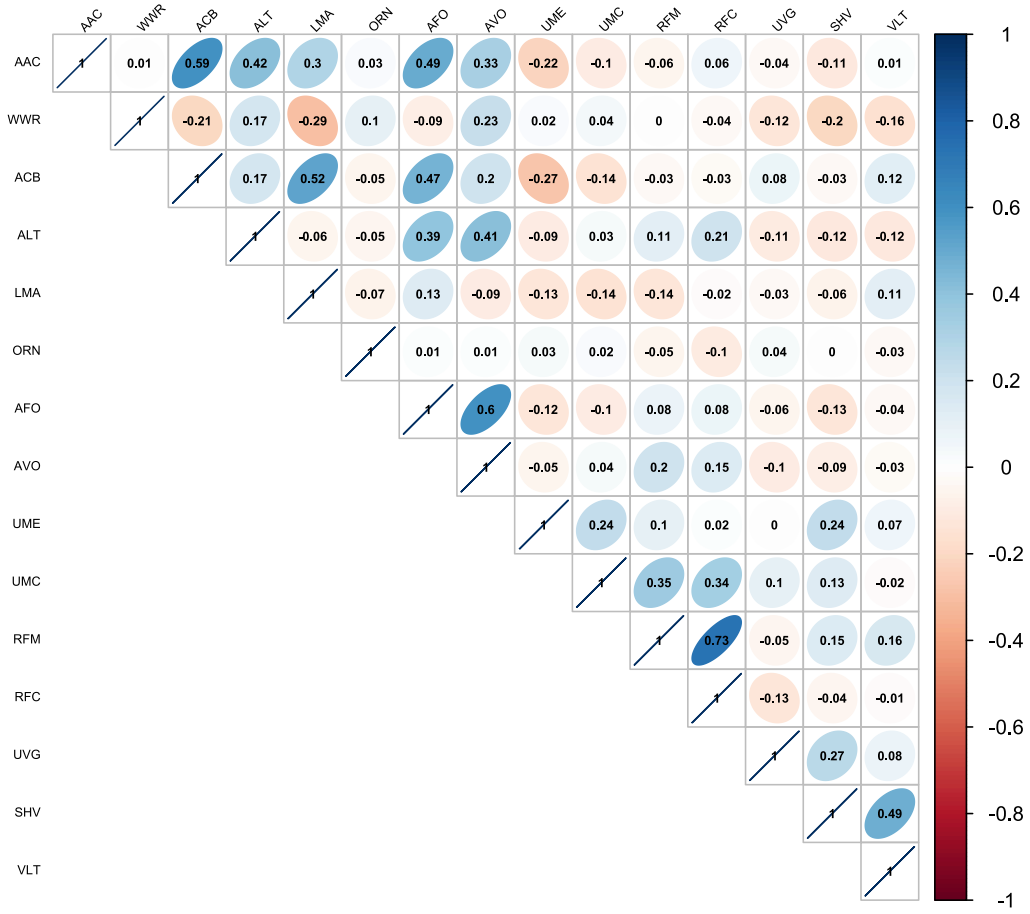

Figura 19: Análisis de correlación de Kendall sin redundancias

8.2. Etapa 2: Reducción de Dimensionalidad Mediante PCA Robusto

Con el objetivo de reducir la dimensionalidad del conjunto de datos y capturar la estructura subyacente de las variables numéricas sesgadas y con atípicos, se aplicó un Análisis de Componentes Principales (PCA) robusto.

El PCA robusto se aplicó sobre las variables previamente transformadas mediante Yeo-Johnson y escaladas robustamente con la mediana y el rango intercuartílico (IQR). Se seleccionó un valor de $k = 10$ para retener las diez primeras componentes principales.

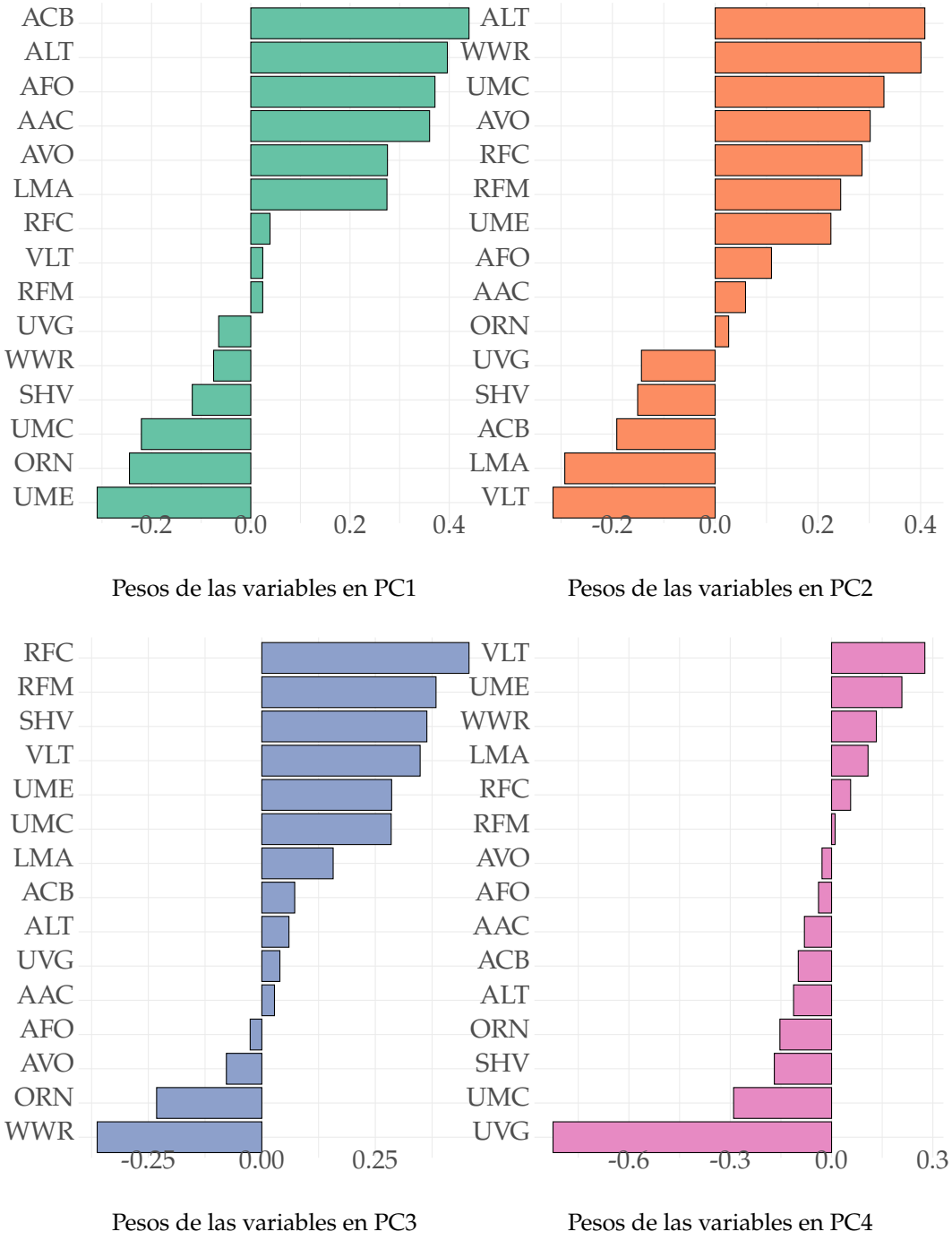

Figura 20: Distribución de pesos de las variables por componente principal

Tabla 14: Cargas de las variables en las 4 primeras componentes principales del PCA robusto

	PC1	PC2	PC3	PC4
AAC	0.36	0.06	0.03	-0.08
WWR	-0.07	0.40	-0.36	0.13
ACB	0.44	-0.19	0.07	-0.10
ALT	0.40	0.41	0.06	-0.11
LMA	0.27	-0.29	0.16	0.11
ORN	-0.24	0.03	-0.23	-0.15
AFO	0.37	0.11	-0.03	-0.04
AVO	0.28	0.30	-0.08	-0.03
UME	-0.31	0.22	0.29	0.21
UMC	-0.22	0.33	0.28	-0.29
RFM	0.02	0.24	0.38	0.01
RFC	0.04	0.29	0.46	0.06
UVG	-0.06	-0.14	0.04	-0.82
SHV	-0.12	-0.15	0.36	-0.17
VLT	0.02	-0.32	0.35	0.28

De acuerdo con la Tabla 14 y la Figura 20, las variables con mayor contribución en cada componente principal son las siguientes: en la PC1 destacan ACB, ALT, AFO, AAC y UME; en la PC2, las variables más relevantes son ALT, WWR, UMC, VLT y LMA; en la PC3, se identifican como principales RFC, RFM, SHV, VLT, WWR y ORN; y en la PC4, las variables con mayor peso son UVG, UMC, VLT y UME.

Tabla 15: Resumen de las varianzas de las 10 primeras componentes principales

Comp Ppal	**Desv.Estándar**	**Varianza Explicada (%)**	**Varianza Acumulada (%)**
PC1	2,08	22,21	22,21
PC2	1,77	17,27	39,48
PC3	1,58	16,59	56,07
PC4	1,29	10,56	66,63
PC5	1,17	8,23	74,86
PC6	1,07	6,35	81,21
PC7	0,86	5,40	86,61
PC8	0,73	3,40	90,01
PC9	0,64	2,63	92,64
PC10	0,49	2,44	95,08

De acuerdo con la Tabla 15, el PC1 explica el 22,21 % de la varianza y PC2 añade un 17,27 %, de modo que ambos concentran el 39,48 %. Al incorporar la tercera componente (PC3), la varianza explicada acumulada asciende al 56,07 %,

y con la cuarta (PC4) se alcanza el 67 %. A partir de este punto la ganancia marginal disminuye: las diez primeras componentes abarcan el 95.08 % de la varianza total.

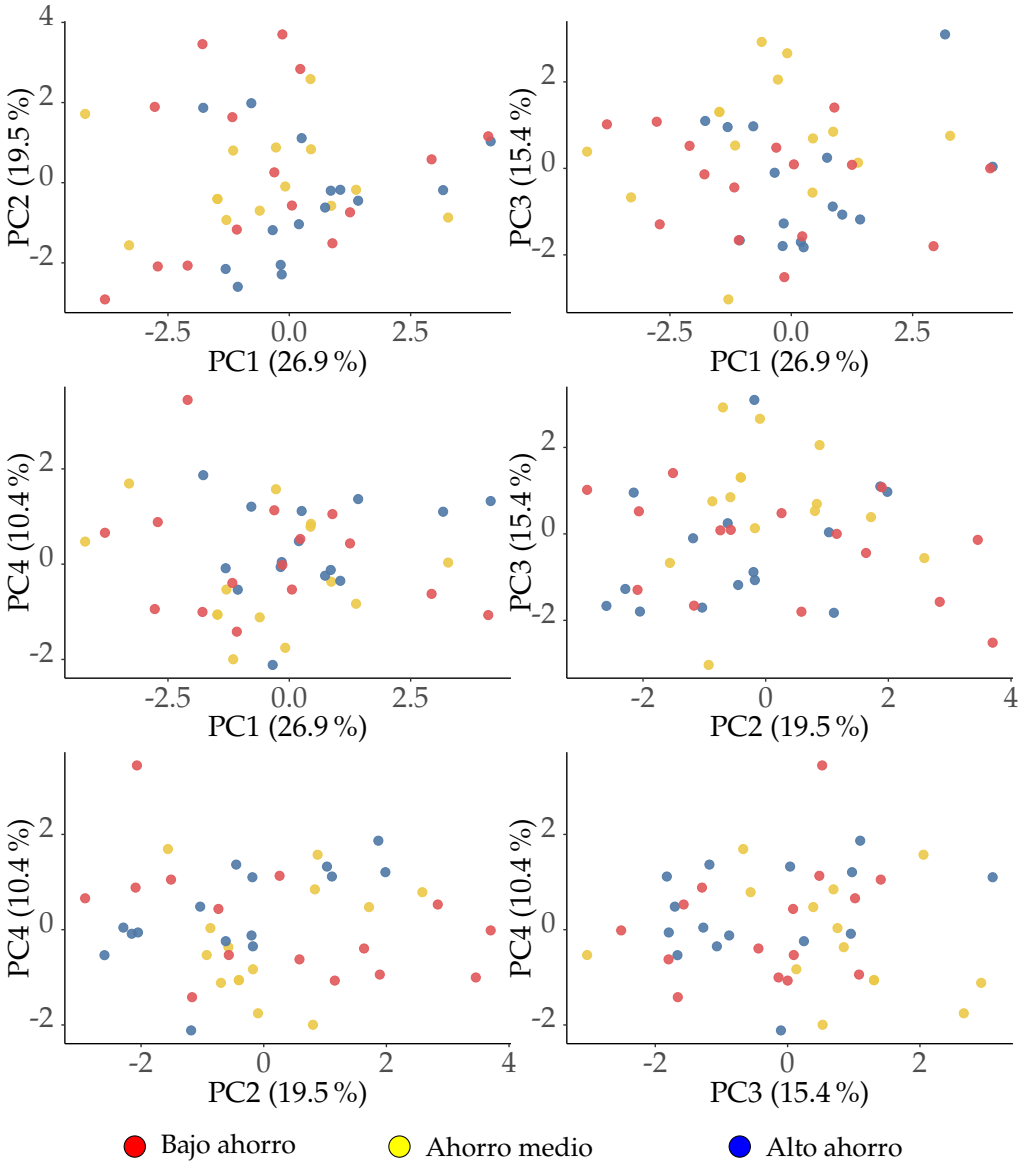

Figura 21: Biplots entre componentes principales PC1–PC4 según niveles de ahorro energético.

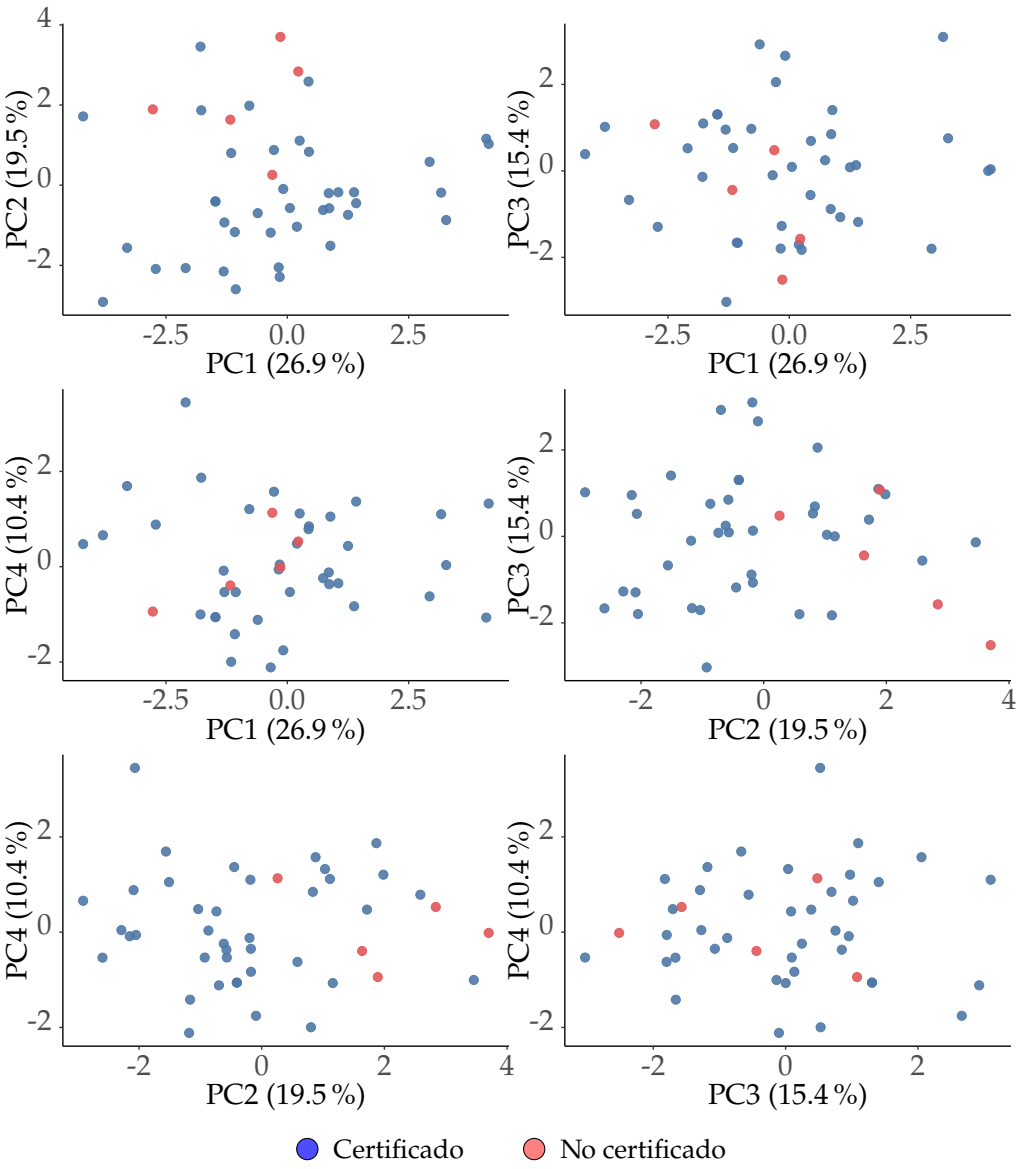

Figura 22: Biplots entre componentes principales PC1–PC4 según estado de certificación.

Las Figuras 21 y 22, muestran los diagramas de dispersión con diferentes combinaciones de componentes principales 1 a 4. De acuerdo con los gráficos de la Figura 21 no se aprecia una separación clara entre los proyectos con niveles bajos, medios y altos de ahorro de energía, mientras que en los gráficos PC2 vs. PC1, PC2 vs. PC3 y PC2 vs. PC4, se aprecia que la componente 2 separa la mayoría de los proyectos entre proyectos certificados y no certificados.

Que la segunda componente principal distinga entre proyectos certificados y no certificados sugiere que las variables con mayor peso en esta dimensión

están fuertemente asociadas al cumplimiento de criterios de eficiencia energética. Entre ellas destacan la altura del edificio (ALT), la relación ventana-muro (WWR), la transmitancia térmica de la cubierta (UMC), el área de ventilación operable (AVO) y las propiedades térmicas y reflectantes de la envolvente (RFC, RFM, UME), todas con cargas positivas. Esto indica que valores altos en estas variables contribuyen significativamente a una mayor puntuación en PC2, y por tanto, están vinculados a proyectos certificados.

En contraste, variables como la transmitancia luminosa visible (VLT), la iluminación artificial (LMA) y el sombreado en ventanas (SHV), con cargas negativas, parecen tener menor incidencia positiva en los resultados energéticos.

Esta diferenciación sugiere que la PC2 está fuertemente influenciada por características físicas del diseño pasivo del edificio, como su geometría, la envolvente térmica y las propiedades constructivas, más que por elementos operacionales. En otras palabras, la certificación de sostenibilidad parece estar más asociada con decisiones arquitectónicas iniciales y con la calidad térmica de los materiales, lo que evidencia la importancia del diseño bioclimático desde etapas tempranas. Este hallazgo coincide con estudios que enfatizan el impacto de la morfología del edificio en su desempeño energético y en la posibilidad de alcanzar estándares internacionales de sostenibilidad.

8.3. Etapa 3: Modelado Mediante Regresión Beta

1. **Usando únicamente las componentes principales obtenidas del PCA robusto**

   El Modelo 1 fue construido utilizando únicamente las componentes principales (PCs), obtenidas mediante un PCA robusto aplicado a las variables cuantitativas previamente transformadas mediante Yeo-Johnson y escaladas con base en el rango intercuartílico (IQR). El objetivo de esta estrategia fue sintetizar la información contenida en los predictores numéricos, reducir la colinealidad y facilitar la interpretación. En esta especificación no se incluyeron variables cualitativas.

   Se implementó un modelo de regresión beta considerando tres funciones de enlace: *logit*, *probit* y *complemento log−log*, empleando validación cruzada estratificada de 5 folds sobre los 44 proyectos de la base de entrenamiento. El menor error absoluto medio (MAE) en la validación cruzada se obtuvo con el enlace *probit* (MAE = 0,0934).

   Usando las 4 componentes, en los tres modelos ajustados (uno por cada enlace), las componentes PC2 y PC3 resultaron estadísticamente significa-

tivas al nivel del 1 %, mientras que PC1 y PC4 no mostraron significancia (valor$-$p >0,47 en todos los casos).

No obstante, tanto el análisis de residuos como la prueba de homocedasticidad indicaron un incumplimiento del supuesto de varianza constante (valor$-$p <0,03), lo que compromete la validez de los resultados del modelo de regresión beta basado exclusivamente en las cuatro componentes principales.

Adicionalmente, se evaluó un modelo que incorporaba tanto la componente 2 como la componente 3 del PCA con los tres enlaces (Ver Tabla 16); ambas resultaron estadísticamente significativa, los resultados se con el enlace *probit* se pueden ver en la Tabla 17.

Tabla 16: Métricas de ajuste del modelo con componentes principales PC2 y PC3

Enlace	**LogLik**	**AIC**	**BIC**	**Pseudo-R^2**
Logit	41,3097	-74,6193	-67,4826	0,2927
Probit	41,4738	-74,9475	-67,8107	0,2735
Clog$-$log	41,1986	-74,3971	-67,2604	0,3004

Tabla 17: Coeficientes del modelo con enlace *probit* (componentes principales PC2 y PC3)

Variable	**Coeficiente**	**valor$-$p**
Intercepto	$-0{,}80658$	$< 2 \cdot 10^{-16}$
PC2	$-0{,}09098$	$0{,}01268$
PC3	$-0{,}12414$	$0{,}00088$

Al modelar cada componente por separado, la componente 2 ofreció un mejor ajuste. El modelo que considera únicamente esta segunda componente principal (PC2) mostró el mejor desempeño al emplear el enlace *clog$-$log*, de acuerdo con las cuatro métricas evaluadas: log-verosimilitud, AIC, BIC y pseudo-R^2 (ver Tabla 18). La Tabla 19 muestra la significancia de la componente. No obstante, dicho modelo no cumple con el supuesto de homocedasticidad (varianza constante de los errores), lo cual limita la validez de sus inferencias.

Tabla 18: Métricas de ajuste de los modelos usando solo PC2 (comparación de funciones de enlace)

Enlace	**LogLik**	**AIC**	**BIC**	**Pseudo**$-R^2$
Logit	35,6261	-65,2522	-59,8964	0,0897
Probit	35,5817	-65,1634	-59,8108	0,0884
Clog$-$*log*	35,6586	-65,3171	-59,9646	0,0901

Tabla 19: Coeficientes significativos del modelo con enlace *probit* usando solo PC2

Variable	**Coeficiente**	**valor**$-$**p**
Intercepto	$-1{,}45548$	$< 2 \cdot 10^{-16}$
PC2 (PCA robusto)	$-0{,}16683$	$0{,}00787$

2. **Combinando las componentes principales con variables categóricas transformadas en variables dummy (sin categoría de referencia).**

   El Modelo 2 amplía la especificación anterior al incorporar, además de las cuatro componentes principales, las variables categóricas representadas mediante variables dummy, sin incluir una categoría de referencia, con el fin de evitar colinealidad estructural en el diseño del modelo.

   Entre los tres enlaces evaluados en la regresión beta (*logit*, *probit* y *complemento log*$-$*log*), el enlace *probit* nuevamente obtuvo el menor error absoluto medio (MAE = 0,0975) en la validación cruzada estratificada de 5 folds.

   En los tres modelos ajustados (uno por cada enlace), y utilizando los 44 proyectos de la base de entrenamiento, la componente principal PC3 resultó estadísticamente significativa al 1 % (valor$-$ p $<$ 0,01). En contraste, ninguna de las variables categóricas representadas como dummy mostró significancia estadística al nivel del 5 %.

   El análisis de residuos como la prueba formal de homocedasticidad evidenciaron violación del supuesto de varianza constante. En consecuencia, los resultados de este modelo no son tenidos en cuenta, ya que el incumplimiento de los supuestos fundamentales compromete la validez de las inferencias.

   Resultados similares se obtuvieron al considerar las variables dummies y solo las componentes 2 y 3.

3. **Utilizando directamente las variables cuantitativas transformadas (sin PCA) junto con las variables dummy derivadas de las variables cate-**

góricas, también sin categoría de referencia.

a) El Modelo 3 parte del conjunto inicial de variables cuantitativas transformadas (sin PCA) y todas las variables categóricas codificadas como dummies, sin categoría de referencia. Para mejorar la parsimonia sin sacrificar capacidad predictiva se:

- eliminaron predictores con varianza casi nula,
- descartaron variables altamente colineales.

Con validación estratificada de 5 folds (44 proyectos) el enlace probit logró el menor MAE (0,1579). Los ajustes finales sobre la muestra completa de entrenamiento se resumen en la Tabla 20. De acuerdo con las métricas, excepto BIC, el enlace que muestra mejores resultados es el probit.

Tabla 20: Métricas de ajuste del Modelo 3 según el enlace

Enlace	**LogLik**	**AIC**	**BIC**	**Pseudo**$-R^2$
Logit	61,12	-68,24	-25,31	0,7011
Probit	62,76	-69,51	-19,56	0,7215
Clog$-$*log*	60,94	-66,86	-22,59	0,6889

Los predictores del modelo regresión beta con enlace probit con valor$-$p <0,05 se resumen en la Tabla 21. Los signos positivos/negativos indican la dirección del efecto sobre el porcentaje de ahorro energético.

Tabla 21: Coeficientes significativos del Modelo 3 con enlace *probit*

Tipo	**Variable**	**Coef.**	**valor$-$p**
Numérica	UVG	$-0,4555$	0,0202
	SHV	$+0336$	0,0274
	VLT	$-0{,}6336$	$< 0{,}001$
Dummy	TIP Residencial	$-0{,}9009$	0,0095
	CIU Bogotá	$+{,}8934$	0,0124

De acuerdo con la la Tabla 21:

- Los proyectos residenciales presentan porcentaje de ahorro energético inferiores a los de otras tipologías, manteniéndose constantes las demás dummies.

- La localización en Bogotá se asocia con mayores niveles de ahorro, manteniéndose constantes las demás dummies para cada caso.
- Entre los predictores continuos, UVG y VLT influyen negativamente, mientras que SHV lo hacen positivamente.

El pseudo$-R^2$ supera el 0,70 en los tres enlaces y alcanza 0,7215 con probit, mejorando sustancialmente al modelo basado sólo en componentes principales.

La supresión de los predictores con varianza casi nula (p. ej. TIP Colegio, TIP Healthcare, CIU Cartagena, ZCL 1B) incrementó la estabilidad del modelo y redujo el riesgo de sobreajuste.

Las pruebas de homocedasticidad e independencia de residuos no mostraron violaciones significativas al 5 %, respaldando la validez del modelo de regresión beta con enlace *probit*.

b) El Modelo 4 descarta por completo los predictores numéricos y se enfoca exclusivamente en variables cualitativas codificadas como dummies *(sin categoría de referencia)* para capturar diferencias estructurales atribuibles a tipología del proyecto, ubicación geográfica y zona de cobertura legal (ZCL). De este modo se evalúa el poder explicativo de las dimensiones categóricas por sí solas.

Con validación estratificada de 5 folds (44 proyectos) la regresión beta con enlace probit obtuvo el menor MAE (0,1007). Los ajustes finales sobre la muestra completa de entrenamiento se resumen en la Tabla 22; el enlace probit presenta el AIC más bajo, pese a que el pseudo-R^2 es ligeramente inferior al de clog–log.

Tabla 22: Métricas de ajuste del Modelo 4 (sólo dummies)

Enlace	**LogLik**	**AIC**	**BIC**	**Pseudo**$-R^2$
Logit	49,05	-77,54	-57,92	0,5690
Probit	50,58	-79,15	-59,52	0,5527
Clog$-$*log*	49,15	-76,29	-56,67	0,5707

Al 5 % de significancia, las variables de la Tabla 23 resultaron relevantes.

Tabla 23: Coeficientes significativos del Modelo 4 con enlace *probit*

Variable dummy	Coeficiente	valor$-$p
AAC	$+0{,}233$	0,0035
AFO	$-0{,}229$	$< 0{,}001$
UMC	$-0{,}172$	0,0183
VLT	$-0{,}246$	0,0076
TIP Residencial	$-0{,}667$	$< 0{,}001$
CIU Bogotá	$+0{,}412$	0,0061

De acuerdo con la Tabla 23 y manteniéndose constantes las demás dummies para cada caso:

- La presencia de la categoría TIP_Residencial reduce el porcentaje de ahorro energético.
- Los proyectos ubicados en Bogotá tienden a presentar ahorros mayores.
- Dummies asociadas a características constructivas o de diseño (AAC, AFO, UMC, VLT) muestran efectos estadísticamente significativos, incluso sin información numérica adicional. En presencia de AAC hay mayor ahorro de energía.

Pese a trabajar exclusivamente con información cualitativa, el Modelo 4 alcanza un desempeño razonable (MAE aproximado de 0,096 y pseudo-R^2 aproximado de 0,57).

La evidencia empírica sugiere que las dimensiones categóricas por sí solas contienen señal predictiva, aunque la exclusión de variables numéricas limita la variabilidad explicada en comparación con el Modelo 3.

Las pruebas de residuos no revelaron violaciones graves de normalidad, homocedasticidad ni independencia al 5 %, por lo que las inferencias del enlace probit son estadísticamente confiables dentro del marco asumido.

c) El Modelo 5 parte del Modelo 3 y elimina todos los predictores que no alcanzaron significancia estadística, con el objetivo de potenciar la *parsimonia* sin sacrificar capacidad predictiva.

La estimación se realizó con los mismos 44 proyectos y validación cruzada estratificada de 5 folds. El enlace *probit* volvió a registrar el menor MAE (0,084), mejorando sensiblemente el desempeño respecto a modelos anteriores.

Los resultados sobre la muestra completa se resumen en la Tabla 24. El enlace *probit* presenta la mejor log-verosimilitud y los valores mínimos de AIC y BIC, mientras que *clog–log* obtiene el pseudo-R^2 más alto (aunque la diferencia es marginal).

Tabla 24: Métricas de ajuste del Modelo 5 (predictores significativos)

Enlace	**LogLik**	**AIC**	**BIC**	**Pseudo**$-R^2$
Logit	47,92	-81,84	-69,35	0,5044
Probit	48,47	-82,93	-70,44	0,4918
Clog$-$*log*	47,47	-80,94	-68,45	0,5065

A un nivel de 5 % las variables de la Tabla 25 se mantienen relevantes. Todos los signos concuerdan con la expectativa sustantiva derivada de modelos previos.

Tabla 25: Coeficientes significativos del Modelo 5 con enlace *probit*

Variable	**Coeficiente**	**valor-p**
AAC	$+0{,}2938$	0,0140
AFO	$-0{,}3306$	$< 0{,}001$
UMC	$-0{,}2216$	0,0400
VLT	$-0{,}2083$	0,0370
TIP Residencial	$-1{,}0244$	0,0006

- TIP_Residencial muestra el mayor impacto absoluto (coef. $-1{,}02$): los proyectos residenciales presentan porcentajes de ahorro energético significativamente menores.
- AAC incrementa el ahorro, mientras que AFO, UMC y VLT lo reducen, en concordancia con su señal física o de diseño.

El Modelo 5 logra el MAE más bajo (0,084) de toda la serie, con una configuración sustancialmente más simple.

Las pruebas de residuos no detectaron violaciones sustantivas de normalidad, homocedasticidad o independencia al 5 % (ver la Figura 23), lo que respalda la validez inferencial del enlace *probit*.

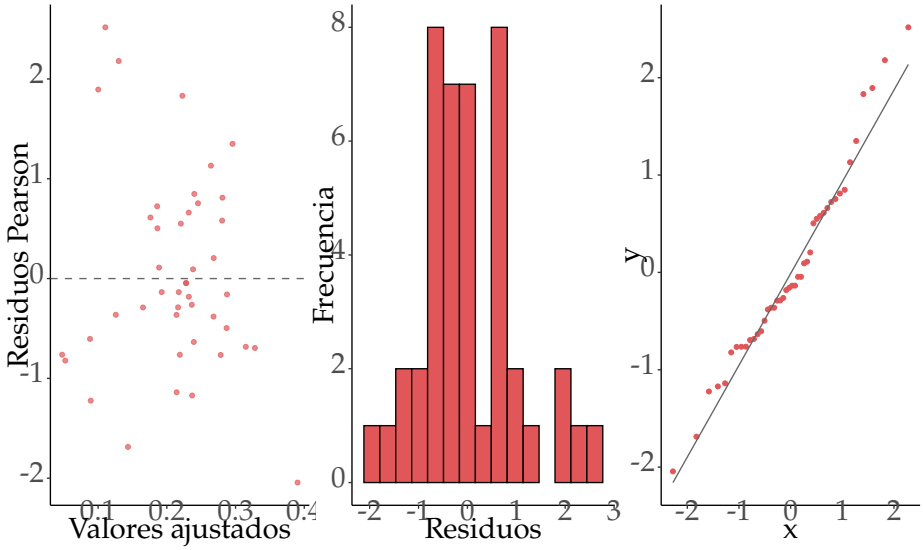

Figura 23: Residuos Modelo 5)

El Modelo 5 presentó el mejor desempeño predictivo global en términos de MAE, sin perder interpretabilidad. La omisión del PCA no perjudicó el ajuste; por el contrario, la combinación directa de variables numéricas estandarizadas con dummies significativas ofreció un modelo parsimonioso. Se confirma que las variables originales contienen relaciones suficientemente fuertes con la variable respuesta, especialmente AFO y TIP. Este modelo representa una primera aproximación a la solución del problema.

Los tres modelos seleccionados como más destacados corresponden, de manera consistente, a los tres mejores ajustes obtenidos con el enlace *probit* dentro de los cinco modelos estimados. La Tabla 26 resume los indicadores clave de cada especificación y muestra por qué los modelos 3, 4 y 5 se consideran los de mayor interés analítico.

Tabla 26: Comparación de desempeño para los cinco modelos con enlace *probit*

Modelo	**LogLik**	**AIC**	**BIC**	R^2
3 (probit)	62,757	-69,513	-19,556	0,7215
4 (probit)	50,575	-79,151	-59,524	0,5527
5 (probit)	48,466	-82,933	-70,443	0,4918
2 (probit)	49,681	-65,362	-35,031	0,5077
1 (probit)	41,932	-71,864	-61,159	0,2801

- *Modelo 3* exhibe el mayor pseudo-R^2 (0,72) y una log-verosimilitud sobresaliente, confirmando su capacidad explicativa al integrar variables cuantitativas y cualitativas.
- *Modelo 4* demuestra que las variables puramente categóricas retienen poder predictivo razonable, aunque con menor R^2.
- *Modelo 5*, pese a su estructura parsimoniosa, alcanza los valores mínimos de AIC y BIC, evidenciando la mejor compensación entre ajuste y complejidad.

En conjunto, estos resultados respaldan la elección de los modelos 3, 4 y 5 como los más útiles para la interpretación y predicción del porcentaje de ahorro energético dentro de la muestra analizada.

En estos modelos, algunas de las variables más influyentes coinciden con las identificadas previamente en la *PC2 del PCA robusto*, que diferenciaba proyectos certificados de no certificados. Entre ellas se destacan: *ALT* (altura del edificio), *WWR* (relación ventana–muro) y *UMC* (transmitancia térmica de la cubierta), que corresponden principalmente a características *físicas* de la edificación. Asimismo, variables como *LMA* (iluminación artificial) y *VLT* (transmitancia luminosa visible) se relacionan con el diseño y desempeño *operacional* del edificio.

Esta coincidencia sugiere que los factores estructurales y de envolvente térmica, junto con características de iluminación y transparencia, son determinantes tanto para la clasificación de certificación como para la predicción del ahorro energético, reforzando la coherencia entre los resultados del PCA y de la regresión beta.

8.4. Etapa 4: Clasificación mediante Análisis Discriminante (RDA)

1. **Certificado y no certificado**

   En el caso con solo componentes del PCA robusto como predictoras, los hiperparámetros óptimos fueron $\gamma^* = 0{,}37$ y $\lambda^* = 0{,}43$. El umbral global se fijó en $t^* = 0{,}887$ minimizando $2\,\mathrm{FN} + 1\,\mathrm{FP}$ sujeto a $\mathrm{Spec} \geq 0{,}50$. Con ese corte (alto), el desempeño fuera de muestra por LOOCV es: $\mathrm{Sens} = 0{,}698$, $\mathrm{Spec} = 0{,}500$, $\mathrm{BalAcc} = 0{,}599$ y $F_{\sqrt{2}} = 0{,}732$ (Tabla 27). Nótese que los valores de sensibilidad muy altos observados durante la búsqueda interna (con el umbral por defecto $t = 0{,}5$) no son directamente comparables con el resultado final: al elevar el umbral a $0{,}887$, la sensibilidad disminuye y la especificidad aumenta hasta cumplir la restricción del 50 %.

   Cuando se combinan componentes del PCA robusto con variables indicadoras (dummies), los hiperparámetros óptimos fueron $\gamma^* = 0{,}34$ y $\lambda^* = 0{,}91$, y el umbral global quedó en $t^* = 0{,}878$ (misma función de costo y restricción). Con ese corte, el LOOCV arroja $\mathrm{Sens} = 0{,}814$, $\mathrm{Spec} = 0{,}500$, $\mathrm{BalAcc} = 0{,}657$ y $F_{\sqrt{2}} = 0{,}833$ (Tabla 27). De nuevo, las cifras de la validación interna reportadas al umbral por defecto ($t = 0{,}5$; p. ej., sensibilidades cercanas a 0.98 junto con especificidades $\approx 0{,}18$) no deben mezclarse con las métricas finales al t^*: un corte alto prioriza la precisión y la especificidad, a costa de la sensibilidad.

Tabla 27: Métricas finales — LOOCV con umbral óptimo t^* para el modelo RDA (certificación), comparando: solo componentes del PCA robusto vs. componentes del PCA robusto más dummies

Modelo	Sens	Spec	BalAcc	$F_{\sqrt{2}}$
PCA robusto	0,698	0,500	0,599	0,732
PCA robusto + dummies	0,814	0,500	0,657	0,833

Al incorporar variables indicadoras junto con las componentes principales en el modelo RDA (certificación), el desempeño fuera de muestra mejora manteniendo la restricción de especificidad: la sensibilidad pasa de 0,698 a 0,814 (con $\mathrm{Spec} = 0{,}500$ en ambos casos), la precisión equilibrada (BalAcc) sube de 0,599 a 0,657 y $F_{\sqrt{2}}$ de 0,732 a 0,833. Esto indica que las dummies aportan señal discriminante adicional sin alterar la prioridad operativa del esquema de costos 2:1, que favorece reducir falsos negativos. Véase la Tabla 27.

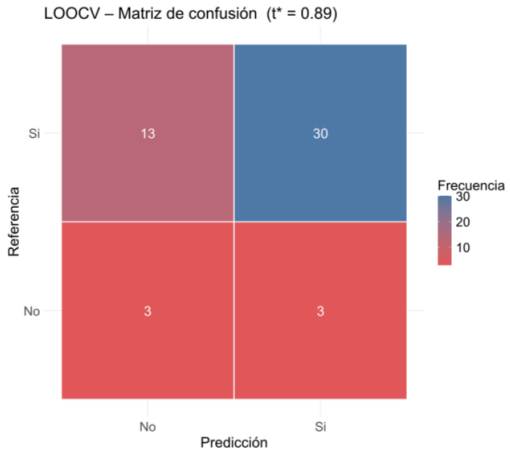

Figura 24: Matriz de confusión, LOOCV (49 pliegues leave-one-out), para el modelo RDA (certificación) usando las componentesprincipales como predictoras

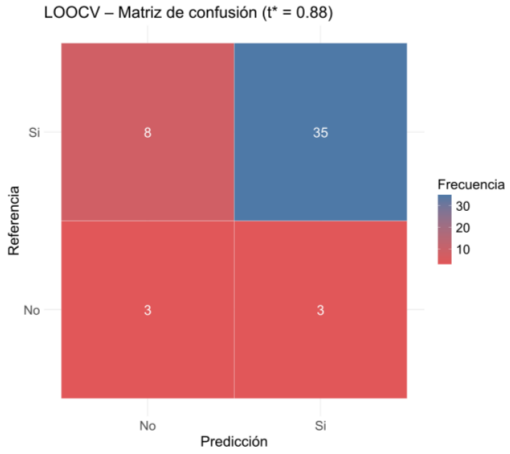

Figura 25: Matriz de confusión (LOOCV, 49 pliegues) para el modelo RDA (certificación) que usa tanto las componentes principales como las variables indicadoras (dummies).

Las matrices de confusión de las Figuras 24 y 25 (LOOCV, $t^* \approx 0{,}88$) muestran el patrón esperado bajo la política de costos 2:1: el umbral fija la especificidad en el borde exigido ($\mathrm{Spec} = 0{,}50$) y empuja al clasificador a recuperar casos «Sí». Con solo componentes se obtienen 30 verdaderos positivos y 13 falsos negativos (con 3 verdaderos negativos y 3 falsos posi-

tivos); al añadir dummies los falsos positivos se mantienen en 3 y los falsos negativos bajan a 8. Esto eleva la sensibilidad de $0{,}698$ a $0{,}814$, mejora la precisión equilibrada y reduce el costo esperado (de 29 a 19 unidades en la función $2 \cdot \mathrm{FN} + \mathrm{FP}$). En términos prácticos, el modelo prefiere «auditar de más» antes que dejar escapar proyectos certificables, que es exactamente la prioridad buscada. Dado que en LOOCV hay muy pocos negativos (6 casos), la especificidad es inestable (un caso cambia Spec en $\approx 0{,}17$), por lo que conviene realizar más estudios al respecto.

Las Figuras 26, 27 y 28 presentan los resultados del modelo RDA (certificación) usando como predictoras únicamente las componentes del PCA robusto. En la curva Precisión–Recall se obtiene AUPR = 0,866; sin embargo, en PR la referencia es la prevalencia de la clase positiva. Dado que «Certificado = Sí» representa 45 de 49 casos ($\approx 0{,}918$), ese 0,866 queda por debajo del piso y no evidencia capacidad de ordenamiento por encima de la estrategia trivial. El ROC–AUC = 0,589 apenas supera el azar, lo que confirma una separación débil entre clases. Esta lectura es coherente: con una clase positiva tan dominante, la PR puede parecer alta en términos absolutos, pero no respecto a la base; la ROC revela la escasa señal del modelo.

En la importancia de variables (Figura 28) la señal discriminante se concentra en PC3 y PC4, no en PC1. Todas las importancias son cercanas ($\approx 0{,}55$–$0{,}66$), sin un componente claramente dominante, lo que sugiere información útil dispersa y débil, en línea con el ROC–AUC $\approx 0{,}59$.

En el escenario binario, usando componentes principales y dummies, el modelo RDA obtiene AUPR = 0,881 y ROC–AUC = 0,643 (Figuras 29 y 30). Frente al modelo con solo PCs hay una mejora leve (AUPR +0,015; ROC–AUC +0,054). Sin embargo, como la clase «Sí» representa 45 de 49 casos ($\approx 0{,}918$), la referencia natural de la PR es muy alta. Por ello, 0,881 queda por debajo del piso y el ranking por probabilidad de «Sí» no supera al azar (ganancia sobre la base $=$ AUPR $-$ prevalencia $= -0{,}037$). La ROC–AUC de 0,643 indica señal existente pero de separación moderada.

En la importancia de variables (Figura 31), PC3 concentra la mayor información discriminante y PC2 también aporta; PC1 no aparece entre las principales, lo que sugiere que la mayor varianza total no coincide con la mejor separación de clases. Entre las categóricas destacan las dummies de tipología de uso (Residencial, Oficinas, Retail), ciudad (Bogotá, Cali) y zona climática (4C, 3C–3A, 2A), indicando segmentación por estos factores en la probabilidad de certificación. Para una interpretación sustantiva

conviene añadir el signo de los coeficientes discriminantes para establecer qué categorías aumentan o disminuyen la probabilidad.

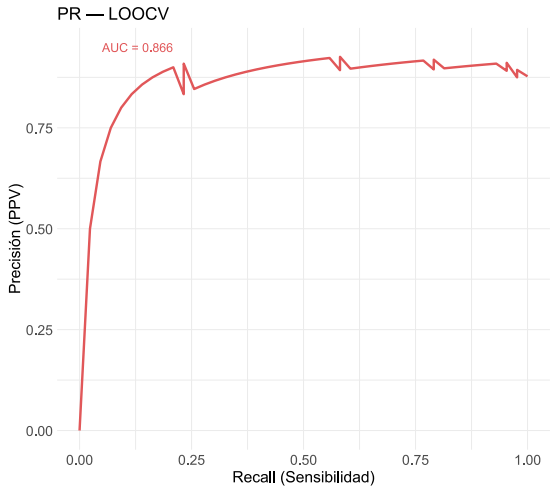

Figura 26: Curva AUC, LOOCV (49 pliegues leave-one-out), del modelo RDA (certificación) utilizando componentes principales

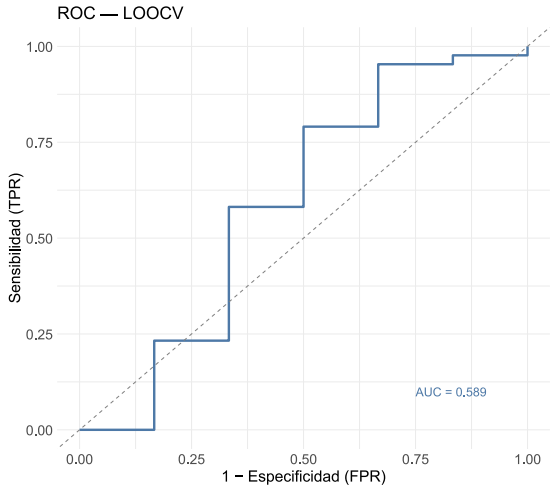

Figura 27: Curva ROC, LOOCV (49 pliegues leave-one-out), del modelo RDA (certificación) utilizando componentes principales

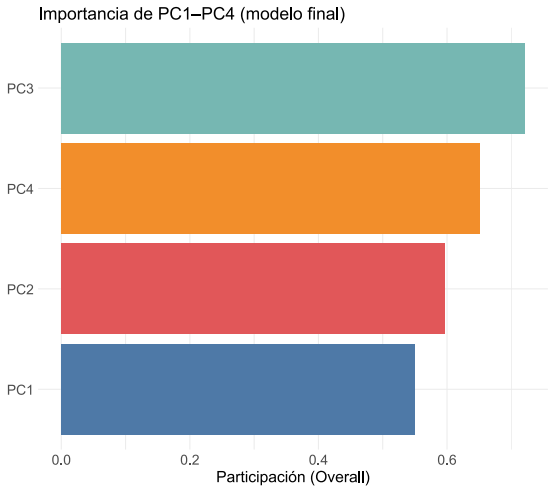

Figura 28: Importancia de las componentes principales usando solamente componentes principales en el modelo RDA (certificación)

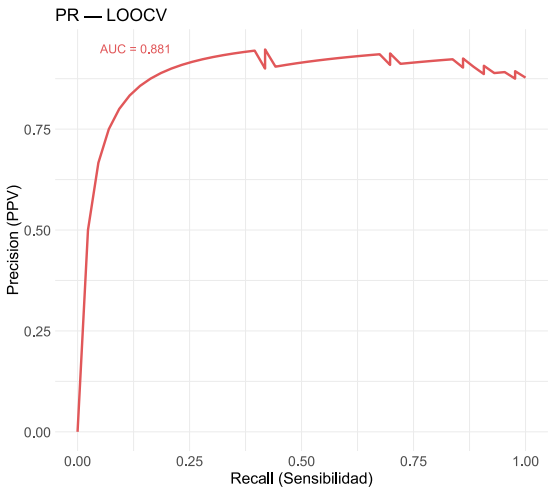

Figura 29: Curva AUC, LOOCV (49 pliegues leave-one-out), del modelo RDA (certificación) utilizando componentes principales y dummies

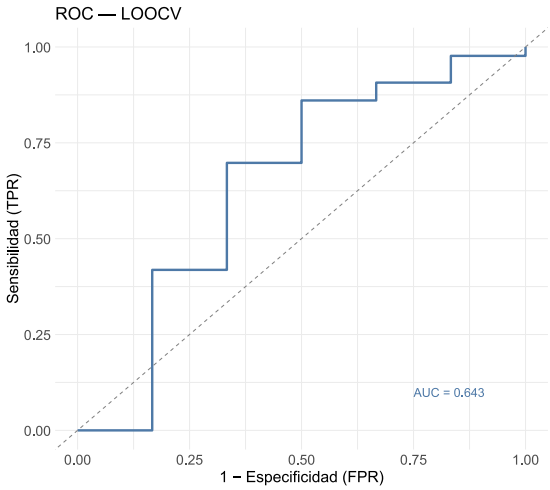

Figura 30: Curva ROC, LOOCV (49 pliegues leave-one-out), del modelo RDA (certificación) utilizando componentes principales y dummies

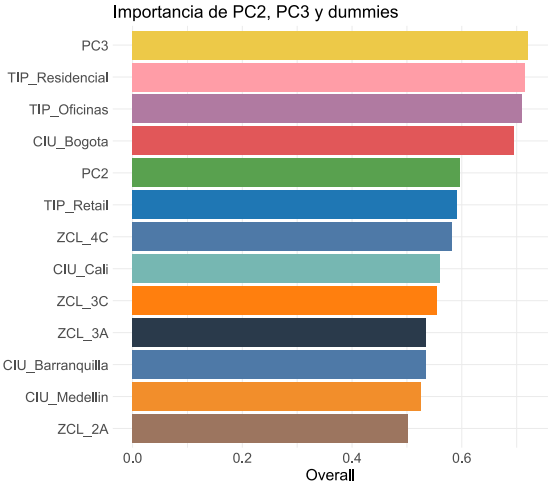

Figura 31: Importancia de las componentes principales y dummies del modelo RDA (certificación)

En el escenario con $k = 2$ categorías de clasificación, las variables más relevantes según el *RDA* fueron *PC3* y *PC4*, ambas derivadas del PCA robusto y asociadas principalmente a características físicas de la envolvente

del edificio, como coeficientes de transmitancia térmica y propiedades de ventilación. Adicionalmente, se identificaron variables categóricas como *TIP Residencial* (proyecto con uso residencial) y *TIP Oficinas* (proyecto con uso de oficinas), que corresponden a características *operacionales* y de uso del inmueble.

Comparando con el PCA, se observa que las variables físicas que componen *PC3* y *PC4*, incluyendo propiedades térmicas de ventanas, muros y cubiertas, también tienen presencia en la regresión beta, lo que indica una consistencia metodológica en su relevancia. Las variables de uso (*TIP Residencial* y *TIP Oficinas*) no emergen en el PCA o la regresión beta con la misma fuerza, lo que sugiere que el RDA captura efectos operacionales que no se reflejan en la reducción de dimensionalidad inicial.

2. **Nivel de ahorro**

   Para el primer caso, modelo RDA (nivel de ahorros), usando solo como variables predictivas las componentes, el proceso de calibración descrito en la sección de metodología arrojó los siguientes *hiperparámetros óptimos* $\gamma^* = 0{,}10$ y $\lambda^* = 0{,}37$. El *umbral de decisión* fue de $t^* = 0{,}563$.

   Para el segundo caso, modelo RDA (nivel de ahorros), usando solo como variables predictivas las componentes, el proceso de calibración determinó los siguientes *hiperparámetros óptimos* $\gamma^* = 0{,}10$ y $\lambda^* = 0{,}49$. El *umbral de decisión* fue de $t^* = 0{,}807$.

   La Tabla 28 resume las macro-métricas (sensibilidad, especificidad, *balanced accuracy* y $F_{\sqrt{2}}$) bajo la relación de costo $2{:}1$ para el modelo RDA del nivel de ahorro. Al incorporar variables indicadoras junto con las componentes principales se observa una mejora pequeña pero consistente en el conjunto de entrenamiento: la sensibilidad macro pasa de $0{,}637$ a $0{,}659$, la *balanced accuracy* de $0{,}727$ a $0{,}744$ ($+0{,}017$) y $F_{\sqrt{2}}$ de $0{,}636$ a $0{,}666$ ($+0{,}030$). Esto sugiere que la información contenida en las dummies aporta discriminación adicional entre las clases (en particular para los niveles de interés operativo, Medio/Alto) sin alterar la prioridad impuesta por la estructura de costos en $k = 3$ (penalización mayor para la subestimación que para la sobreestimación).

   En el *hold-out* de 5 proyectos, las métricas perfectas ($1{,}000$ en todas) deben interpretarse con cautela debido al tamaño muestral muy reducido, que puede producir estimaciones inestables. En consecuencia, el énfasis analítico recae en las macro-métricas del entrenamiento con validación cruzada y en la coherencia con la función de costo definida.

Tabla 28: Macro-métricas con relación de costo 2:1: comparación entre usar sólo componentes principales y componentes principales más dummies. *Entrenamiento* con 44 proyectos y *Prueba* con 4 proyectos

Modelo	Conjunto	Sens	Spec	BalAcc	$F_{\sqrt{2}}$
PCA robusto	Entrenamiento	0,637	0,817	0,727	0,636
	Prueba	1,000	1,000	1,000	1,000
PCA robusto + dummies	Entrenamiento	0,659	0,828	0,744	0,666
	Prueba	1,000	1,000	1,000	1,000

Para el escenario con $k = 3$ categorías, las variables con mayor importancia fueron *TIP Oficinas*, junto con las componentes principales *PC2*, *PC3* y *PC4*. La *PC2*, previamente identificada en el PCA como un discriminador clave entre proyectos certificados y no certificados, agrupa principalmente variables físicas como la altura del edificio (*ALT*), la relación ventana–muro (*WWR*), la transmitancia térmica de la cubierta (*UMC*), el área de ventilación operable (*AVO*) y propiedades térmicas y reflectantes de la envolvente (*RFC*, *RFM*, *UME*). Por su parte, *PC3* y *PC4* mantienen la relevancia observada en el escenario $k = 2$, mientras que *TIP Oficinas* incorpora el componente operacional.

En la comparación entre métodos, se confirma que variables físicas como *ALT*, *WWR* y *UMC* tienen un papel determinante en la eficiencia energética estimada, mientras que el uso del edificio (*TIP Oficinas*) añade una dimensión operacional que potencia la capacidad discriminativa del RDA. Esto indica que la combinación de características físicas y operacionales ofrece una representación más completa del comportamiento energético de los proyectos.

8.5. Etapa 5: Predicciones en Conjunto de Prueba

1. **Modelado Mediante Regresión Beta**

   La Tabla 29 presenta el error absoluto medio (MAE) obtenido al aplicar los Modelos 3, 4 y 5 (todos con enlace *probit*) al conjunto de prueba compuesto por 5 edificios. El Modelo 5 resulta nuevamente el más preciso:

Tabla 29: Error absoluto medio (MAE) en el conjunto de prueba

Modelo	MAE
3–probit	0,2963
4–probit	0,2583
5–probit	0,2188

Para los tres modelos el MAE de prueba es mayor que el reportado en validación cruzada, lo que evidencia una *brecha de generalización*. Aunque esperada en cierto grado, esta diferencia sugiere que los modelos aún capturan parte del ruido no generalizable del entrenamiento. La brecha es menor en el Modelo 5, lo cual refuerza la conveniencia de su estructura parsimoniosa.

Entre las implicaciones y líneas de mejora para predecir el porcentaje de ahorro de energía se pueden explorar como trabajos futuros:

a) *Regularización adicional*. Explorar *ridge* o *elastic-net* sobre el predictor lineal podría reducir la varianza de los coeficientes y, con ello, la brecha entre entrenamiento y prueba.

b) *Validación más robusta*. Con sólo 49 edificios, validar mediante *Leave One Out* (LOO) o *repeated k-fold* proporcionaría estimaciones más estables de la incertidumbre.

c) *Aumento de datos*. Incorporar edificios adicionales ya sea de nuevas mediciones o registros históricos permitiría un testeo externo más fiable y disminuiría la varianza del MAE.

d) *Análisis de data*. Es recomendable revisar si los 5 edificios de prueba presentan combinaciones inusuales de tipología, ubicación o ZCL; de ser así, se debe considerar ampliar la cobertura de dichos patrones en el entrenamiento.

El Modelo 5–*probit* presenta mejor desempeño tanto en validación cruzada como en prueba externa, pero el valor mayor MAE en el conjunto de prueba indica que el modelo aún no generaliza plenamente. Es necesario seguir explorando soluciones como reforzar la regularización, emplear métodos de validación más exigentes y, cuando sea posible, ampliar la base de datos para consolidar la capacidad predictiva.

2. **Clasificación mediante Análisis Discriminante (RDA)**

Las Figuras 32, 33 y 34 reflejan los resultados del modelo RDA para predecir nivel de ahorro de energía usando como variables predictivas las componentes principales.

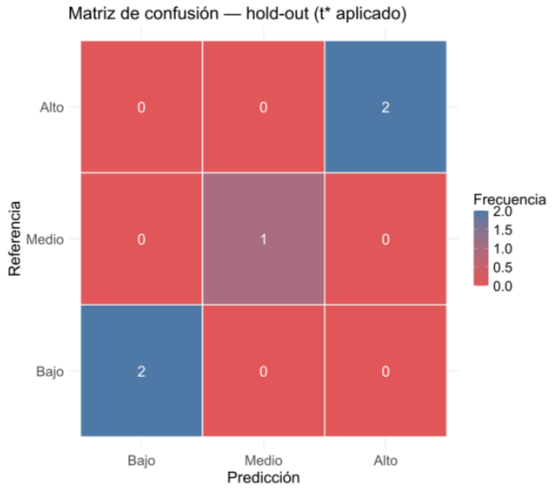

Figura 32: Matriz de confusión del modelo que predice nivel de ahorro usando como predictivas las componentes principales (conjunto de prueba- 5 proyectos)

La matriz de confusión, predice perfecto en prueba (2 Bajo, 1 Medio, 2 Alto correctos). Eso explica que las macro-métricas del hold-out salgan en 1,0 (Figura 32). Las curvas ROC por clase muestran AUC(Bajo)=1,0, AUC(Medio)=1,0, AUC(Alto)=0,833 (Figura 33). En la Figura 34 se aprecia que las componentes PC2 y PC4 son las más importantes, luego aparece PC3 y, por último, PC1. Es decir, la capacidad de separar clases no está en PC1, aunque sea la que explica más varianza total.

Las Figuras 35, 36 y 37 muestran los resultados del modelo RDA para predecir nivel de ahorro de energía usando como variables predictivas las componentes principales y las variables dummies.

En el hold-out de 5 proyectos la matriz de confusión queda perfecta: dos "Alto", dos "Bajo" y uno "Medio" clasificados correctamente (Figura 35). Las ROC por clase reflejan lo mismo: "Bajo" y "Medio" aparecen prácticamente perfectas y "Alto" muy alta (Figura 36).

La importancia de variables muestra un patrón claro: el tipo de edificio pesa mucho (destaca "Oficinas", seguido de "Residencial" y "Retail"); después vienen las componentes PC2, PC4 y PC3, y algo más abajo, PC1. Es decir, la mayor varianza no es la que mejor separa niveles de ahorro. También aparecen con peso medio la ciudad (resalta Barranquilla; Bogotá, Cali y Medellín algo menos) y la zona climática (4C, 3C, 2A), lo que sugiere

segmentación por uso y clima/localidad además de los patrones resumidos por las PCs (Figura 37). Añadir dummies aporta señal e interpretabilidad para distinguir Bajo/Medio/Alto.

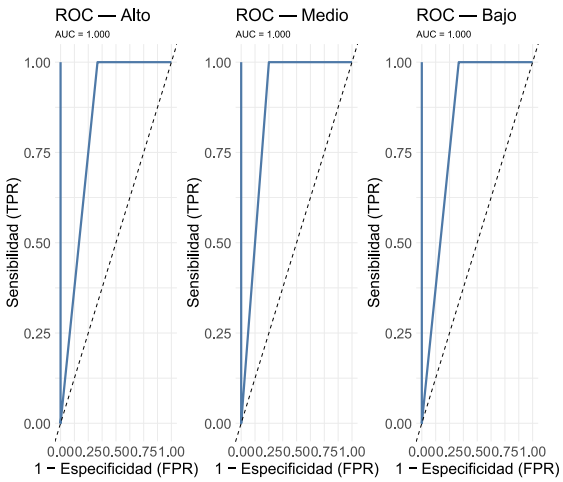

Figura 33: Curvas ROC del modelo que predice nivel de ahorro usando como predictivas las componentes principales (conjunto de prueba- 5 proyectos)

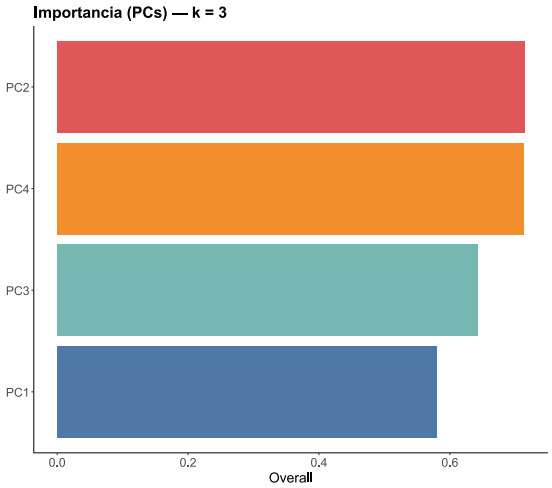

Figura 34: Importancia de las variables del modelo que predice nivel de ahorro usando como predictivas las componentes principales (conjunto de prueba- 5 proyectos)

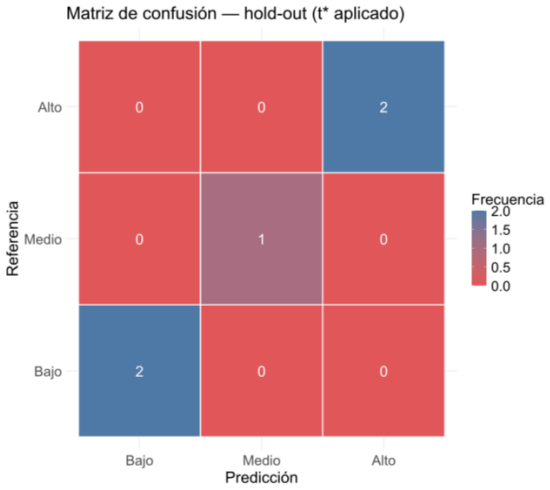

Figura 35: Matriz de confusión del modelo que predice nivel de ahorro usando como predictivas las componentes principales más las dummies (conjunto de prueba- 5 proyectos)

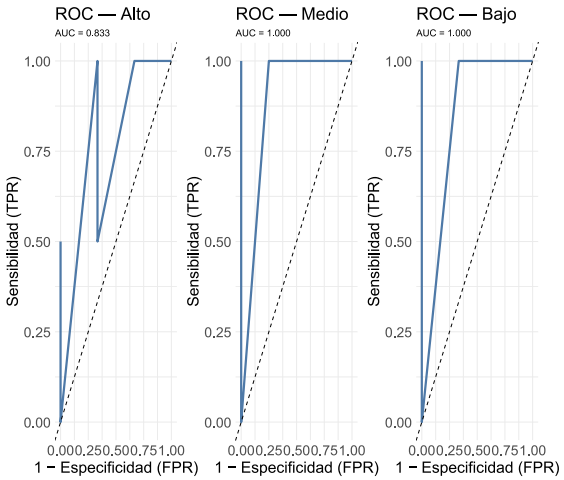

Figura 36: Curvas ROC del modelo que predice nivel de ahorro usando como predictivas las componentes principales más las dummies (conjunto de prueba- 5 proyectos)

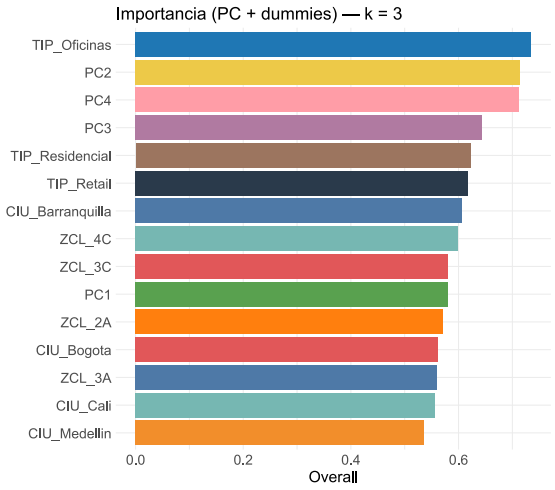

Figura 37: Importancia de las variables del modelo que predice nivel de ahorro usando como predictivas las componentes principales más las dummies (conjunto de prueba- 5 proyectos)

9 Discusión

9.0.1. Análisis de modelamiento

1. Los resultados confirman la idoneidad de la *regresión beta* como marco para modelar el porcentaje de ahorro energético, al estar naturalmente acotado en $(0, 1)$ y admitir funciones de enlace flexibles. El mejor desempeño correspondió al Modelo 5—*probit*, que, tras depurar los predictores a los verdaderamente significativos, alcanzó el menor MAE tanto en validación cruzada interna como en la prueba externa. Esta parsimonia sugiere que gran parte de la señal explicativa se concentra en un subconjunto reducido de variables (p. ej. *AAC*, *AFO* y la dummy *TIP Residencial*), mientras que la eliminación de covariables redundantes mitigó el sobreajuste. No obstante, la brecha observada entre el error de validación y el de prueba, aunque esperable con una muestra de prueba de apenas cinco edificios, evidencia cierta captura de ruido no explicado propio del entrenamiento. Esto resalta la necesidad de incorporar técnicas de regularización (ridge o elastic-net) y de ampliar la base de mediciones para robustecer la generalización.

2. El análisis de componentes principales (PCA) robusto se adoptó para condensar la alta dimensionalidad de las variables cuantitativas y mitigar la influencia de atípicos. Sin embargo, los diagramas de dispersión de las cuatro primeras componentes no revelaron patrones claros que permitiesen discriminar *a priori* entre edificios con ahorro bajo, medio o alto. Aunque la componente 2 si logró diferenciar entre inmuebles certificados y no certificados. Consecuentemente, cuando dichas cuatro componentes se emplearon como únicas covariables en la regresión beta, el ajuste resultante fue inferior al de los modelos que combinaron variables originales y dummies categóricas: el pseudo-R^2 cayó a 0,28 y el MAE se incrementó, evidenciando que la compresión lineal sacrificó información predictiva

relevante.

Estos hallazgos resaltan que la naturaleza lineal del PCA puede esconder estructuras no lineales ligadas al ahorro energético. Como trabajo futuro se sugiere explorar técnicas de reducción supervisada con mayor número de edificios, que podrían presentar mejor los límites a los distintos niveles de desempeño energético y a la certificación.

3. En el escenario binario «certificación sí/no», el RDA con componentes principales más dummies supera de forma consistente al modelo con solo componentes. Con validación LOOCV y el umbral global optimizado bajo el esquema de costos 2:1 (imponiendo $\mathrm{Spec} \geq 0{,}50$), el modelo final ($\gamma^* = 0{,}34$, $\lambda^* = 0{,}91$, $t^* = 0{,}878$) alcanza $\mathrm{Sens} = 0{,}814$, $\mathrm{Spec} = 0{,}500$, $\mathrm{BalAcc} = 0{,}657$ y $F_{\sqrt{2}} = 0{,}833$. En comparación, el modelo con solo componentes obtiene $\mathrm{Sens} = 0{,}698$, $\mathrm{BalAcc} = 0{,}599$ y $F_{\sqrt{2}} = 0{,}732$. Operativamente, los falsos negativos disminuyen de 13 a 8 manteniendo los falsos positivos en 3, por lo que el costo esperado pasa de $2 \cdot 13 + 3 = 29$ a $2 \cdot 8 + 3 = 19$. La capacidad de separación global sigue siendo moderada ($\mathrm{AUC}_{\mathrm{ROC}} \approx 0{,}64$), pero el desempeño al umbral t^* es coherente con la prioridad de negocio de evitar dejar escapar proyectos certificables aun aceptando auditorías adicionales.

   Debe considerarse que la clase «Sí» es muy mayoritaria ($\pi \in [0{,}88, 0{,}92]$), de modo que la $\mathrm{AUC}_{\mathrm{PR}}$ se debe leer frente a su línea base π: un valor alto en términos absolutos puede no representar ganancia real si no supera con claridad ese piso. En estas condiciones, resulta más informativo reportar las métricas calculadas al umbral operativo t^* elegido con la función de costos 2:1. Por ello, se recomienda presentar sensibilidad, especificidad, valor predictivo positivo y el costo esperado al t^*, junto con la matriz de confusión LOOCV. Este enfoque está alineado con la decisión de negocio (reducir falsos negativos) y evita conclusiones engañosas derivadas del desbalance extremo, donde $\mathrm{AUC}_{\mathrm{ROC}}$ puede ser moderada y la $\mathrm{AUC}_{\mathrm{PR}}$ aparentar buen desempeño sin superar la referencia π.

   Además, con muy pocos negativos en LOOCV (se observan 6 casos negativos), la estimación de la especificidad es inestable: un solo caso modifica Spec en $\approx 0{,}17$. Por tanto, para estudios futuros es recomendable acompañar las métricas de intervalos de confianza binomiales y de una evaluación de calibración (curva de confiabilidad y/o Brier score) que refleje la fiabilidad de las probabilidades. También es útil reportar la estabilidad del umbral t^* y de las métricas a través de los remuestreos internos (distribución de t_i^* y de $F_{\sqrt{2}}$).

En cuanto a la implementación práctica, debe tenerse en cuenta que t^* y el valor predictivo positivo dependen tanto de la prevalencia como de los costos: si la prevalencia operativa difiere de la observada, se requiere reoptimizar el umbral bajo la misma función de costos o, alternativamente, presentar una curva de beneficio neto (decision–curve) que muestre la utilidad esperada a distintos cortes. Interpretar el modelo al t^* y bajo el esquema de costos proporciona una medida más honesta de impacto que el uso aislado de áreas bajo curva en un contexto de fuerte desbalance.

En términos de interpretación, la señal se concentra en PC3 y PC2, junto con dummies de tipología de uso, ciudad y zona climática, lo que sugiere segmentación por estos factores en la probabilidad de certificación. Dado que la importancia de variables no informa sobre el signo del efecto, es conveniente complementar con los coeficientes discriminantes o efectos marginales para establecer qué categorías aumentan o disminuyen la probabilidad estimada.

4. En k=3 (Bajo/Medio/Alto), el modelo RDA que combina componentes principales con dummies ofrece una mejora pequeña pero consistente frente al modelo con solo componentes: en la validación cruzada del entrenamiento aumentan la sensibilidad macro, la balanced accuracy y el F-beta, lo que indica que las variables categóricas aportan señal útil sin contradecir la prioridad de penalizar más las subestimaciones.

   En el conjunto de prueba de cinco proyectos el desempeño es perfecto, pero ese tamaño de muestra es tan reducido que cualquier error movería las métricas alrededor de veinte puntos; por tanto, ese resultado debe interpretarse con cautela.

   La importancia de variables muestra un patrón claro: pesan especialmente el tipo de edificio (destacan Oficinas, luego Residencial y Retail), la ciudad (resalta Barranquilla) y la zona climática (4C, 3C, 2A), junto con las componentes PC2 y PC4; en cambio, PC1, pese a explicar más varianza total, discrimina menos entre niveles de ahorro.

   En términos prácticos, el modelo permite ordenar proyectos por nivel de ahorro priorizando no subestimar los casos Medio/Alto y aporta criterios interpretables para la toma de decisiones (uso, clima y rasgos de diseño).

   Para afianzar estas conclusiones, en estudios futuros conviene repetir la evaluación con particiones externas más amplias o con bootstrap, calibrar probabilidades y ajustar el corte operativo según la prevalencia y los costos reales. Adicionalmente, no considerar la categorización por terciles para

determinar niveles de ahorro y en su lugar, establecer valores clave de la certificación, por ejemplo 5 % o 2,5 % de ahorro energético.

9.0.2. Protocolo

Es necesario un protocolo que guíe a los constructores en la construcción de edificaciones sostenibles y que les permita obtener la certificación ASHRAE. Proponemos que las siguientes etapas como protocolo:

1. *Definición del alcance del proyecto*: Establecer claramente los objetivos del análisis energético, tales como la reducción del consumo, la mejora de la eficiencia y el cumplimiento normativo. También, seleccionar las edificaciones a evaluar, considerando su uso, tamaño y características específicas.

2. *Recolección de datos climáticos y de consumo energético*: Recopilar datos meteorológicos relevantes para la ubicación del edificio, incluyendo temperatura, humedad y radiación solar, así como el consumo energético actual del edificio, desglosado por tipo de uso (iluminación, climatización, equipos eléctricos).

3. *Modelado energético*: Utilizar herramientas de simulación energética para modelar el comportamiento energético del edificio conforme a los estándares *ASHRAE*. Este modelado debe considerar variables como los materiales de construcción, los sistemas de climatización y los patrones de ocupación.

4. *Análisis de cargas térmicas*: Aplicar las metodologías *ASHRAE* para calcular las cargas térmicas del edificio, tomando en cuenta factores como la radiación solar y el confort térmico (*ASHRAE* 55). Comparar los resultados obtenidos con un edificio de referencia, según las normas *ASHRAE*.

5. *Simulación y predicción de escenarios*: Crear diferentes escenarios que simulen la implementación de medidas de eficiencia energética (como mejoras en el aislamiento o la implementación de sistemas HVAC más eficientes). Posteriormente, comparar los resultados obtenidos de la modelación enérgetica con datos reales, para validar la precisión de los modelos. Seleccionar tres modelos preliminares para definir cuál es el más preciso y realizar ajustes adicionales según sea necesario.

6. *Elaboración de un informe y plan de implementación*: Elaborar un informe detallado que incluya los hallazgos, análisis y recomendaciones basadas

en los resultados obtenidos. Proponer además un plan para implementar las medidas recomendadas, priorizando aquellas que presenten un mayor impacto en el ahorro energético.

9.0.3. Recomendaciones

A continuación proponemos una serie de recomendaciones para la planificación y ejecución de un plan que priorice las medidas con mayor impacto en la eficiencia energética. Los pasos detallados para esta etapa podrían incluir las siguientes actividades:

- *Entender los objetivos de la metodología ASHRAE*: Consiste en verificar y documentar que las instalaciones de un edificio cumplen con los requerimientos del propietario y los criterios de diseño, garantizando una operación eficiente y confiable desde el primer día.
- *Aplicar los estándares ASHRAE relevantes*: Utilizar la norma *ASHRAE* 90.1 para eficiencia energética mínima en edificios y la *ASHRAE* 189.1 para edificios de alto rendimiento.
- *Seguir el proceso de comisionamiento ASHRAE*: Implementar las fases del proceso de comisionamiento de *ASHRAE*, que incluyen la planificación previa al diseño, el desarrollo del plan de comisionamiento, la aplicación de los estándares, el diagnóstico y la mejora continua.
- *Utilizar los métodos de cálculo de carga térmica ASHRAE*: Emplear los métodos de cálculo de carga térmica recomendados por *ASHRAE* para predecir el consumo energético del edificio, considerando las condiciones climáticas locales.
- *Implementar el modelo de predicción energética*: Desarrollar un modelo de predicción energética que utilice los métodos de cálculo de carga térmica *ASHRAE* y considere factores como la temporada, la ocupación, las condiciones climáticas y las características constructivas del edificio. Validar el modelo con mediciones reales de consumo energético para asegurar su precisión.
- *Aplicar estrategias de control y optimización*: Utilizar el modelo de predicción energética para implementar estrategias avanzadas de control y optimización del sistema HVAC, con el fin de maximizar el ahorro energético del edificio.

Conclusiones

1. Este estudio propone una metodología basada en técnicas de ciencia de datos para modelar el nivel de ahorro energético y la certificación de edificios, utilizando como base sus características físicas y operativas. Mediante un enfoque estructurado que incluyó un preprocesamiento exhaustivo, reducción de dimensionalidad mediante PCA robusto, modelado con regresión beta y clasificación mediante Análisis Discriminante regularizado (RDA), se demostró que es posible identificar patrones significativos que explican tanto el comportamiento energético como la probabilidad de contar con certificación según estándares como *ASHRAE*.

2. La regresión beta con enlace *probit* se identificó como el enfoque más adecuado para modelar el porcentaje de ahorro energético, al respetar la naturaleza acotada de la variable respuesta y lograr el menor valor de MAE tanto en la validación cruzada interna como en la prueba externa. Además, lo hizo con una especificación parsimoniosa, lo que refuerza su utilidad práctica. No obstante, la diferencia de desempeño entre los conjuntos de entrenamiento y prueba, amplificada por el reducido tamaño de la muestra de prueba, sugiere la necesidad de validar estas conclusiones en estudios con muestras más amplias y validaciones independientes más rigurosas.

3. En el contexto de la regresión beta, los predictores *Área acondicionada del edificio (AAC)*, *Área de fachada occidente (AFO)*, *Transmitancia térmica de la cubierta (UMC)*, *Transmitancia luminosa visible (VLT)* y la variable categórica *TIP_Residencial* —que indica si el edificio es de uso residencial— mostraron una asociación estadística destacada con el ahorro energético. Estas variables representan características tanto físicas (envolvente térmica, geometría y propiedades ópticas) como operacionales (uso del edificio), lo que ofrece una base sólida para orientar intervenciones de diseño y futuras investiga-

ciones en eficiencia energética.

4. Aunque el PCA robusto permitió reducir la dimensionalidad del conjunto de variables y mitigar la influencia de valores atípicos, no logró capturar adecuadamente las estructuras subyacentes asociadas al nivel de ahorro energético. Sin embargo, en el estatus de certificación la componente 2 logró una separación de proyectos certificados y no certificados. Su uso exclusivo como conjunto de predictores en la regresión beta resultó en un desempeño subóptimo del modelo, evidenciado por un MAE elevado.

5. Para el modelo RDA, como criterio general para ambos escenarios (certificación y nivel de ahorro energético), la selección de métricas y la fijación del umbral se alinean explícitamente con el costo relativo de los errores. En certificación (k=2) el error más caro es el falso negativo (dejar sin certificar a quien sí cumple), por lo que conviene optimizar un F-beta con mayor peso a la sensibilidad, elegir el umbral minimizando el costo esperado bajo esa asimetría y reportar desempeño al corte operativo (sensibilidad, especificidad, valor predictivo positivo y costo), más que apoyarse en áreas bajo curva susceptibles a sesgos por prevalencia. En nivel de ahorro (k=3) la penalización debe distinguir entre subestimación (más costosa) y sobreestimación (menos costosa), por lo que se recomienda usar un F-beta macro-promediado que priorice la sensibilidad por clase, evaluar con métricas macro (sensibilidad, especificidad y balanced accuracy) y cuantificar el costo multiclase a partir de la matriz de confusión completa, diferenciando explícitamente los errores por sub- y sobre-clasificación. En ambos casos, es importante calibrar probabilidades, verificar la estabilidad del umbral y de las métricas en remuestreos, contrastar la utilidad con curvas de beneficio neto y reoptimizar el umbral cuando cambien la prevalencia o la estructura de costos; solo así la metodología del RDA puede reflejar de manera adecuada las prioridades operativas y el impacto real de los errores.

6. En nivel de ahorro (k=3), con clases balanceadas, añadir dummies aporta una ganancia modesta del modelo RDA pero robusta y suma interpretabilidad. El uso del edificio, la ciudad y la zona climática, junto con las componentes más relevantes, ayudan a separar los niveles y a orientar decisiones de mejora. El resultado perfecto observado en el conjunto de prueba es indicativo pero no concluyente por su tamaño reducido. En certificación (k=2) se observa el mismo patrón, al pasar de solo componentes a componentes más dummies aumenta la sensibilidad y la balanced accuracy, se reducen de forma clara los falsos negativos sin incrementar los

falsos positivos y desciende el costo esperado bajo la política 2:1; aunque la separación global sigue siendo moderada y las áreas bajo curva deben leerse con cautela por la alta prevalencia de "sí", el desempeño medido al umbral operativo refleja mejor la prioridad de negocio. En conjunto, el RDA que integra componentes y dummies, con umbral y evaluación alineados al costo de los errores, es la opción recomendada para ambos objetivos.

7. El análisis mostró que variables relacionadas con la *envolvente térmica* (RFC, RFM, UMC,UME,UVG), la *(ALT)*, la *relación ventana muro (WWR)*, los *sistemas mecánicos* (ACB, AAC) presentan una fuerte influencia en la eficiencia energética del edificio. Esto sugiere que estas variables pueden ser utilizadas no solo para comprender el comportamiento energético observado, sino también para anticipar la probabilidad de alcanzar una certificación energética antes de que esta sea solicitada formalmente. La consistencia en la aparición de estas variables en los modelos PCA, regresión beta y RDA refuerza su relevancia como indicadores clave en el diseño y evaluación de edificaciones sostenibles.

8. En particular, la integración de modelos predictivos como la regresión beta para variables continuas (porcentaje de ahorro energético) y modelos de clasificación para variables categóricas (certificación sí/no, nivel de ahorro) representa una herramienta valiosa para la toma de decisiones en etapas tempranas del diseño. Este enfoque podría permitir anticipar el desempeño energético potencial y puede servir como base para definir estrategias orientadas a mejorar la eficiencia energética y guiar los proyectos hacia la obtención de certificaciones mediante intervenciones fundamentadas en evidencia.

9. Los resultados de este trabajo están condicionados por el tamaño limitado de la muestra, especialmente en lo que respecta a edificios no certificados. Por ello, una línea futura de mejora consiste en ampliar el conjunto de datos con una muestra más representativa de edificaciones, tanto certificadas como no certificadas, especialmente en la región ecuatorial. Esto permitiría aumentar la generalización de los modelos desarrollados y mejorar su precisión predictiva, reforzando aún más el valor de la ciencia de datos en el contexto de la eficiencia energética y la sostenibilidad en la edificación.

Bibliografía

[1] Interempresas. «Las Bases del Modelado Energético.» Consultado: 22 de noviembre de 2024. (2022), dirección: https://www.interempresas.net/Climatizacion/Articulos/382269-Las-bases-del-modeladoenergetico.htm.

[2] Distrito Energético. «Eficiencia Energética en Edificaciones para el Desarrollo Urbano Sostenible.» Consultado: 22 de noviembre de 2024. (2023), dirección: https://www.distritoenergetico.com/eficiencia-energetica-en-edificaciones-para-el-desarrollo-urbano-sostenible/.

[3] L. Ortega-Díaz, J. Cárdenas-Rangel y G. Osma-Pinto, «Estrategias de Predicción de Consumo Energético en Edificaciones: Una Revisión,» *Revista TecnoLógicas*, vol. 26, n.º 58, e300, 2023. DOI: 10.22430/22565337.2650.

[4] I. Martin y M. Alarcón, «Evolución de las Metodologías para el Cálculo de Cargas Térmicas en Edificaciones, Desarrolladas por la ASHRAE,» en *Memorias de la Asociación Nacional de Energía Solar (ANES)*, Consultado: 22 de noviembre de 2024, Oaxaca, México, 2004, págs. 113-116. dirección: https://cimav.repositorioinstitucional.mx/jspui/bitstream/1004/1073/1/Publicacion%20Congreso%20Nal%20ANES%20Oaxaca%202004%20Metodos%20ASHRAE.pdf.

[5] ASHRAE, *ASHRAE Standard 90.1 - Energy Standard for Buildings Except Low-Rise Residential Buildings*. American Society of Heating, Refrigerating y Air-Conditioning Engineers, 2010.

[6] J. Hernandez y L. Gómez, «Energy performance assessment in tropical climates: challenges for international standards,» *Energy and Buildings*, vol. 219, pág. 110 020, 2020.

[7] M. Zárate y A. Prieto, «Climatic zoning and building energy codes in Colombia: A gap to bridge,» *Sustainability*, vol. 11, n.º 13, pág. 3701, 2019.

[8] D. Cuervo y M. Espinosa, «Barreras para el financiamiento de proyectos sostenibles en América Latina,» *Revista de la CEPAL*, n.º 125, págs. 119-134, 2018.

[9] Consejo Colombiano de Construcción Sostenible, *Avances de la Construcción Sostenible en Colombia*, Disponible en: https://construccionsostenible.org, 2017.

[10] USGBC, *LEED en Colombia*, Disponible en: https://www.usgbc.org/leed, 2020.

[11] F. Bazurto y C. Ramírez, «Adaptación de herramientas de simulación energética para climas tropicales,» *Revista Ingeniería y Competitividad*, vol. 23, n.º 2, e230206, 2021.

[12] U.S. Green Building Council, *LEED v4 for Building Design and Construction Reference Guide*, 2013 Edition. Washington, D.C.: U.S. Green Building Council, 2013, pág. 335, Prerequisite: Minimum Energy Performance.

[13] PNUD. «Guía para la Implementación de Programas Nacionales Voluntarios de Huella de Carbono en América Latina.» Consultado: 22 de noviembre de 2024. (2022), dirección: https://climatepromise.undp.org/sites/default/files/research_report_document/PNUD_Carbono_ESP_v02.pdf.

[14] J. Cárdenas-Rangel, G. Osma-Pinto y G. Ordóñez-Plata, «Herramienta Metodológica para la Evaluación Energética Mediante Simulación de Edificaciones en el Trópico,» *Revista UIS Ingenierías*, vol. 18, n.º 2, págs. 259-268, 2019. DOI: 10.18273/revuin.v18n2-2019024.

[15] Ministerio de Vivienda, Ciudad y Territorio. «Resolución 549 de 2015 con Anexos.» Consultado: 22 de noviembre de 2024. (2015), dirección: https://camacol.co/sites/default/files/Resoluci%C3%B3n%20549%20de%202015%20con%20Anexos.pdf.

[16] U.S. Green Building Council, *LEED Certification Fees*, Accessed: Feb. 16, 2025, 2024. dirección: https://www.usgbc.org/tools/leed-certification/fees.

[17] Organismo Internacional de Energía Atómica. «Reducción de los Gases de Efecto Invernadero.» Consultado: 22 de noviembre de 2024. (2023), dirección: https://www.iaea.org/es/temas/reduccion-de-los-gases-de-efecto-invernadero.

[18] U.S. Green Building Council, *LEED v4 for Building Design and Construction*, Online, https://www.usgbc.org/leed/v4, 2023.

[19] ASHRAE, *Standard 90.1 – Energy Standard for Buildings Except Low-Rise Residential Buildings*. American Society of Heating, Refrigerating y Air-Conditioning Engineers, 2022.

[20] P. Torcellini, M. Deru y D. Crawley, «Performance Metrics Protocols: A Common Language for Commercial Buildings,» National Renewable Energy Laboratory (NREL), inf. téc. NREL/TP-550-38600, 2004. dirección: https://www.nrel.gov/docs/fy04osti/38600.pdf.

[21] W. Tenorio, «Modelos Energéticos de Edificios Urbanos: Una Revisión Sistemática de Literatura,» *Sapienza: International Journal of Interdisciplinary Studies*, vol. 3, n.º 2, págs. 334-345, 2022. DOI: 10.51798/sijis.v3i2.342.

[22] A. Cortes. «Cómo Modelar Eficiencia Energética en Edificios Utilizando Dinámica de Sistemas.» Consultado: 20 de julio de 2024. (2023), dirección: https://www.software-shop.com/index.php/contenido/video/6937.

[23] ASHRAE, *ASHRAE Standard 90.1-2010: Energy Standard for Buildings Except Low-Rise Residential Buildings*. American Society of Heating, Refrigerating y Air-Conditioning Engineers, 2010.

[24] D. B. Crawley et al., «EnergyPlus: creating a new-generation building energy simulation program,» *Energy and Buildings*, vol. 33, n.º 4, págs. 319-331, 2001.

[25] P. Torcellini y D. Crawley, «Building Energy Software Tools Directory,» *National Renewable Energy Laboratory (NREL)*, 2006.

[26] D. B. Crawley, J. W. Hand, M. Kummert y B. T. Griffith, «Contrasting the capabilities of building energy performance simulation programs,» *Building and environment*, vol. 43, n.º 4, págs. 661-673, 2008.

[27] U. G. B. Council, *LEED v4: Building Design and Construction Guide*, https://www.usgbc.org/, Acceso en julio de 2025, 2020.

[28] B. Mercado-Colín y L. Romero-Guzmán, «Simulaciones Energéticas: Herramientas Diagnóstico-Pronóstico para la Evaluación de Edificaciones,» *Revista Legado de Arquitectura y Diseño*, vol. 17, n.º 31, 2022. dirección: https://www.redalyc.org/journal/4779/477970601013/html/.

[29] J. Correa. «Simulación Energética de Edificios.» Consultado: 22 de noviembre de 2024. (2015), dirección: https://www.acrlatinoamerica.com/202003175781/articulos/automatizacion-de-edificios/simulacion-energetica-de-edificios.html.

[30] W. Shen, W. Tang, A. Siripanan et al., «Critical Success Factors in Thailand Green Building Industry,» *Journal of Asian Architecture and Building Engineering*, vol. 16, n.º 2, págs. 317-324, 2017. DOI: 10.3130/jaabe.16.317. dirección: https://doi.org/10.3130/jaabe.16.317.

[31] Banco de Desarrollo de América Latina (CAF). «Reporte de Eficiencia Energética en Colombia.» Consultado: 22 de noviembre de 2024. (2016), dirección: https://scioteca.caf.com/bitstream/handle/123456789/960/Reporte%20EE%20en%20Colombia.pdf?isAllowed=y&sequence=1.

[32] Consejo Colombiano de Construcción Sostenible (CCCS), *Estado de la Construcción Sostenible en Colombia 2022*. Bogotá: CCCS, 2021.

[33] U.S. Green Building Council, *LEED v4 for Building Design and Construction Reference Guide*, 2013 Edition. Washington, D.C.: U.S. Green Building Council, 2013, pág. 405, Credit: Optimize Energy Performance.

[34] I. T. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, 2002.

[35] P. J. Rousseeuw y A. M. Leroy, *Robust Statistics*. Wiley, 2005.

[36] P. Filzmoser, K. Hron y C. Reimann, «Principal component analysis for compositional data with outliers,» *Environmetrics*, vol. 20, n.º 6, págs. 621-632, 2008.

[37] W. N. Venables y B. D. Ripley, *Modern Applied Statistics with S*. New York, NY: Springer-Verlag, 2002.

[38] M. Hubert, P. J. Rousseeuw y T. Verdonck, «Robust PCA for Skewed Data and Its Outlier Map,» *Computational Statistics and Data Analysis*, vol. 53, págs. 2264-2274, 2009. DOI: 10.1016/j.csda.2008.05.027.

[39] J. Fox y S. Weisberg, *An R Companion to Applied Regression*, 3rd. Thousand Oaks, CA: Sage, 2019.

[40] D. Hawkins y S. Weisberg, «Combining the Box-Cox Power and Generalized Log Transformations to Accommodate Nonpositive Responses in Linear and Mixed-Effects Linear Models,» *South African Statistical Journal*, vol. 51, págs. 317-328, 2017.

[41] I.-K. Yeo y R. A. Johnson, «A new family of power transformations to improve normality or symmetry,» *Biometrika*, vol. 87, n.º 4, págs. 954-959, 2000.

[42] P. J. Huber y E. M. Ronchetti, *Robust Statistics*, 2.ª ed. John Wiley & Sons, 2011.

[43] S. L. P. Ferrari y F. Cribari-Neto, «Beta Regression for Modeling Rates and Proportions,» *Journal of Applied Statistics*, vol. 31, n.º 7, págs. 799-815, 2004. DOI: 10.1080/0266476042000214501.

[44] M. Smithson y J. Verkuilen, «A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables,» *Psychological Methods*, vol. 11, n.º 1, págs. 54-71, 2006. DOI: 10.1037/1082-989X.11.1.54.

[45] D. M. Hawkins y G. J. McLachlan, «High-Breakdown Linear Discriminant Analysis,» *Journal of the American Statistical Association*, vol. 92, págs. 136-143, 1997. DOI: 10.2307/2291457.

[46] V. Todorov y A. M. Pires, «Comparative Performance of Several Robust Linear Discriminant Analysis Methods,» *REVSTAT Statistical Journal*, vol. 5, págs. 63-83, 2007. DOI: 10.57805/revstat.v5i1.42.

[47] B. Efron y R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press, 1994.

[48] U.S. Green Building Council, *LEED v4 for Building Design and Construction*. Washington, DC, USA: U.S. Green Building Council, 2013, Referencia oficial de LEED y su relación con ASHRAE 90.1.

[49] A. M. D. Rodríguez, J. B. C. Martínez, J. P. Acción, A. C. Izaguirre y R. G. Álvarez, «Modelo Matemático para Predecir el Consumo de Energía Eléctrica en la Universidad de Cienfuegos,» *Universidad y Sociedad*, vol. 12, n.º 4, págs. 21-31, 2020. dirección: https://rus.ucf.edu.cu/index.php/rus/article/view/1608.

[50] S. Zaragoza-Fernández, J. Tarrío-Saavedra, S. Naya, J. López-Beceiro y A. Álvarez-García, «Estimación del Impacto de Acciones en la Rehabilitación de la Eficiencia Energética en la Edificación Residencial,» *DYNA*, vol. 81, n.º 186, págs. 186-193, 2014. DOI: 10.15446/dyna.v81n186.39930.

[51] A. M. Álvarez, «Análisis y Predicción de Perfiles de Consumo Energético en Edificios Públicos Mediante Técnicas de Minería de Datos,» Tesis doctoral. Consultado: 22 de noviembre de 2024, Tesis doctoral, Universidad de Oviedo, Oviedo, España, 2012. dirección: https://digibuo.uniovi.es/dspace/bitstream/10651/13420/2/Tesis_AntonioMoran.pdf.

[52] A. Acosta, I. Gonzales, J. Zamarreño y V. Álvarez, «Modelo para la Predicción Energética de una Instalación Hotelera,» *Revista Iberoamericana de Automática e Informática Industrial*, vol. 8, n.º 4, pág. 13, 2011. DOI: 10.1016/j.riai.2011.09.001.

[53] The MathWorks Inc., *MATLAB version: 9.13.0 (R2022b)*, En línea, Natick, Massachusetts, United States, 2022. dirección: https://www.mathworks.com.

[54] R Core Team, *R: A Language and Environment for Statistical Computing*, Consultado: 22 de noviembre de 2024, Vienna, Austria: R Foundation for Statistical Computing. dirección: https://www.R-project.org/.

[55] V. Todorov, *rrcov: Scalable Robust Estimators with High Breakdown Point*, Paquete de R versión 1.7-5, 2024. dirección: https://cran.r-project.org/package=rrcov.

[56] J. Fox y S. Weisberg, *An R Companion to Applied Regression*, 3rd. Thousand Oaks, CA: Sage, 2019. dirección: https://www.john-fox.ca/Companion/.

[57] R Core Team, *The R Stats Package*, Paquete de R versión 4.4.1, 2023. dirección: https://CRAN.R-project.org/package=stats.

[58] F. Cribari-Neto y A. Zeileis, «Beta Regression in R,» *Journal of Statistical Software*, vol. 34, n.º 2, págs. 1-24, 2010. DOI: 10.18637/jss.v034.i02.

[59] M. Kuhn, «Building Predictive Models in R Using the caret Package,» *Journal of Statistical Software*, vol. 28, n.º 5, págs. 1-26, 2008. DOI: 10.18637/jss.v028.i05. dirección: https://www.jstatsoft.org/v028/i05.

[60] C. Weihs, U. Ligges, K. Luebke y N. Raabe, «klaR Analyzing German Business Cycles,» en *Data Analysis and Decision Support*, D. Baier, R. Decker y L. Schmidt-Thieme, eds., Berlin: Springer-Verlag, 2005, págs. 335-343.

Agradecimientos

A nuestras familias y amistades, por el apoyo incondicional que sostuvo cada paso. A las personas y las instituciones que financiaron y facilitaron tiempo y recursos esenciales.