Algoritmo basado en técnicas supervisadas para la detección y estimación de la ocupación en aulas de clase del edificio E3T

María Fernanda Rangel Jiménez

Trabajo de Grado para Optar al Título de Ingeniera Electricista

Tatiana Stella Zarate Luna

Trabajo de Grado para Optar al Título de Ingeniera Electrónica

Director

Juan Manuel Rey López

Ph.D. en Ingeniería Electrónica (UPC)

Codirector

Juan Diego Caballero Peña

M.Sc.en Ingeniería Eléctrica

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2026

Agradecimientos

María Fernanda Rangel Jiménez

Primeramente quiero agradecer a Dios, por permitirme cumplir el sueño de ser Ingeniera Electricista de la universidad que siempre soñé, por ser mi fortaleza en cada paso, por darme la salud, la sabiduría y la perseverancia necesarias para llegar hasta este momento.

Agradezco a mis padres, por su amor incondicional y su ejemplo de vida. A mi mamá, Andrea Jiménez, que nunca me dejó rendirme, me sostuvo con sus palabras de aliento en los días difíciles y celebró conmigo cada logro. A mi papá, Jorge Rangel, por su esfuerzo constante y por mantenerme durante todos estos años, brindándome la tranquilidad necesaria para concentrarme en mis estudios y cumplir mis metas.

Agradezco a mi abuela Leonor Jiménez y a mi tía Sonia Ricaurte, quienes me guiaron a lo largo de los años como si fueran unas madres para mí, con amor, sabiduría y dedicación, aconsejándome y acompañándome en todo momento.

A mi hermano David Rangel y a todas mis mascotas, que me motivaron día a día incluso en los momentos en los que no veía progreso, recordándome con su compañía y cariño que siempre vale la pena seguir adelante, fueron mi motor a lo largo de estos años.

A mis compañeros de universidad, quienes compartieron conmigo largas jornadas de estudio, desvelos, alegrías, tristezas y satisfacciones a lo largo de cada semestre. Agradezco a mi pareja Daniel Morales, por acompañarme en este último año, ser un pilar emocional y motivarme a continuar.

A mi director Juan Manuel Rey y codirector Juan Diego Caballero, por su guía, paciencia y confianza en mi capacidad para afrontar este reto académico. Agradezco a mi compañera Tatiana Zarate, porque a pesar discusiones logramos afrontar este desafío juntas y trabajar en equipo. A la Universidad Industrial de Santander, a la Facultad de Ingenierías Fisicomecanicas y a la E3T, por brindarme una formación académica y humana de excelencia. A todos los profesores que, con su enseñanza y ejemplo, dejaron huella en mi camino profesional.

A mi familia en general —primos, tíos y abuelas—, por acompañarme en cada etapa de este camino, por sus palabras de ánimo y por el orgullo sincero con el que siempre han celebrado mis logros. Su apoyo ha sido una fuente permanente de motivación.

Este logro es el resultado de un esfuerzo constante, de enfrentar y superar obstáculos, de aprender a no rendirme frente a la adversidad. Cada reto fue una oportunidad para crecer, y cada caída, una lección para levantarme más fuerte. Hoy, al mirar atrás, comprendo que este camino no solo me ha dado un título, sino también la certeza de que la disciplina, la dedicación y la fe pueden convertir cualquier meta en una realidad.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Resumen

Título: Algoritmo basado en técnicas supervisadas para la detección y estimación de la ocupación en aulas de clase del edificio E3T*

Autoras: María Fernanda Rangel Jiménez y Tatiana Stella Zárate Luna**

Palabras clave: Ocupación, edificaciones institucionales, aprendizaje supervisado, IoT.

Descripción:

Este trabajo de grado aborda la detección y estimación de la ocupación en edificaciones institucionales, información relevante para la operación eficiente de sistemas HVAC y la gestión energética. Se emplearon mediciones *in situ* de variables ambientales y eléctricas, tales como temperatura, humedad, consumo energético, detección de movimiento y estado de puertas y ventanas, recolectadas en dos aulas del edificio de la Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones (E3T) de la Universidad Industrial de Santander. Se desarrollaron y evaluaron tres técnicas de aprendizaje supervisado con potencial para esta aplicación: *Support Vector Machine* (SVM), *Random Forest* (RF) y *Multi-Layer Perceptron* (MLP). Los modelos se entrenaron mediante una partición entre conjuntos de entrenamiento y prueba, y su desempeño se evaluó con métricas de clasificación como *Accuracy*, *Recall*, *Specificity*, AUC y *F1-score*. Los resultados muestran que el modelo RF presenta un desempeño más consistente en la detección y estimación de la ocupación, superando a SVM y MLP en la mayoría de las métricas evaluadas. Estos hallazgos constituyen una referencia metodológica para el desarrollo de sistemas de gestión energética basados en datos en edificaciones institucionales.

**Trabajo de Grado.

***Facultad de Ingenierías Fisicomecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones (E3T). Director: Ph.D. Juan Manuel Rey López. Codirector: M.Sc. Juan Diego Caballero Peña.

Abstract

Title: Algorithm Based on Supervised Learning Techniques for Occupancy Detection and Estimation in Classrooms of the E3T Building*

Authors: María Fernanda Rangel Jiménez and Tatiana Stella Zárate Luna**

Keywords: Occupancy, institutional buildings, supervised learning, IoT.

Description:

This undergraduate thesis addresses occupancy detection and estimation in institutional buildings, information that is relevant for the efficient operation of HVAC systems and energy management. In-situ measurements of environmental and electrical variables were used, including temperature, humidity, energy consumption, motion detection, and door and window status, collected from two classrooms located in the building of the School of Electrical, Electronic and Telecommunications Engineering (E3T) at the Universidad Industrial de Santander. Three supervised learning techniques with potential for this application were developed and evaluated: *Support Vector Machine* (SVM), *Random Forest* (RF), and *Multi-Layer Perceptron* (MLP). The models were trained using a train–test split scheme, and their performance was evaluated using classification metrics such as *Accuracy*, *Recall*, *Specificity*, AUC, and *F1-score*. The results show that the RF model presents more consistent performance in occupancy detection and estimation, outperforming SVM and MLP in most of the evaluated metrics. These findings constitute a methodological reference for the development of data-driven energy management systems in institutional buildings.

**Undergraduate Thesis.

***Faculty of Physical-Mechanical Engineering. School of Electrical, Electronic and Telecommunications Engineering (E3T). Advisor: Ph.D. Juan Manuel Rey López. Co-advisor: M.Sc. Juan Diego Caballero Peña.

Introducción

El sector de las edificaciones y de la construcción es responsable de aproximadamente el 37 % de las emisiones globales de CO_2 y del 34 % del consumo mundial de energía, lo que lo convierte en un componente crítico dentro de los Objetivos de Desarrollo Sostenible (ODS), particularmente los ODS 7 (energía asequible y no contaminante), 11 (ciudades y comunidades sostenibles) y 13 (acción por el clima) (United Nations, 2015; United Nations Environment Programme, 2023). En consecuencia, la eficiencia energética en edificaciones se ha consolidado como una prioridad para entidades gubernamentales y académicas. En Colombia, este enfoque se encuentra respaldado por la Ley 1715 de 2014 y el Decreto 2407 de 2024, los cuales promueven la adopción de tecnologías orientadas a la gestión inteligente de la energía y la integración de sistemas de medición que faciliten la toma de decisiones informadas sobre el consumo energético (Decreto 2407, 2024; Ley 1715, 2014).

Un aspecto determinante en la eficiencia energética de los edificios es el comportamiento de los ocupantes, dado que su presencia influye directamente en la demanda de iluminación, climatización, ventilación y otros servicios. Estudios previos reportan que una detección precisa de la ocupación puede contribuir a reducciones del consumo energético asociadas a sistemas HVAC en un rango entre el 10 % y el 30 % (Candanedo & Feldheim, 2016; Li & Fan, 2020). Sin embargo, alcanzar estimaciones confiables continúa siendo un desafío debido a la variabilidad temporal y espacial de la ocupación, así como a restricciones asociadas con la privacidad, la arquitectura de los edificios y los costos de implementación.

El desarrollo de tecnologías basadas en el Internet de las Cosas (IoT) ha ampliado el conjunto de estrategias disponibles para abordar este problema. Los métodos basados en cámaras ofrecen altos niveles de precisión, pero presentan limitaciones relacionadas con la privacidad de los usuarios y elevados requerimientos computacionales (Chaudhari et al., 2024; Zeleny et al., 2024). Por su parte, las técnicas que utilizan señales WiFi o Bluetooth permiten inferir la presencia mediante dispositivos personales, aunque pueden asociar la identidad de los usuarios con su ubicación (Zhang & Jain, 2019). En contraste, la fusión de sensores ambientales, como concentración

de CO_2, temperatura, humedad y niveles de iluminación, permite estimar la ocupación de manera no intrusiva, con menores costos de implementación y un impacto reducido sobre la privacidad (Chitnis et al., 2025; Mena et al., 2022; Mena-Martinez et al., 2024; Mohammadabadi et al., 2022).

En cuanto al modelado de la ocupación, las técnicas de aprendizaje supervisado han mostrado un desempeño competitivo tanto en tareas de detección como en la estimación de niveles de ocupación. Algoritmos como Random Forest (RF), máquinas de soporte vectorial (SVM) y modelos híbridos basados en redes neuronales convolucionales (CNN) y *XGBoost* han demostrado su capacidad para capturar relaciones lineales y no lineales en los datos, alcanzando resultados favorables en escenarios controlados y reales (Breiman, 2001; J. Chen et al., 2018; Mohammadabadi et al., 2022). No obstante, el desempeño de estos modelos depende de factores como la calidad de los datos, la ubicación de los sensores, la frecuencia de muestreo y la representatividad de las condiciones observadas durante el periodo de medición.

Esta investigación evalúa diversos modelos de aprendizaje supervisado para identificar y cuantificar la presencia de personas en aulas de la Universidad Industrial de Santander (UIS). La relevancia del estudio reside en vincular la precisión de los datos de ocupación con la optimización energética institucional, fundamentando el análisis en registros reales y parámetros de validación estrictos. La metodología planteada integra una revisión sistemática de la literatura y la implementación de modelos de aprendizaje supervisado sobre datos tabulares procedentes de sensores y registros operativos. La base de datos analizada abarca variables como potencia activa, condiciones ambientales (temperatura y humedad), estado de aberturas y detección de movimiento, capturadas con una resolución temporal de cinco minutos, entre el 16 de febrero y el 30 de abril de 2024.

A través de métricas de clasificación binaria y estimación, se realiza un análisis comparativo del desempeño de las técnicas en cuanto a precisión y robustez. Aunque el alcance no incluye la ejecución de estrategias de control, los hallazgos proporcionan el sustento técnico necesario para optimizar la gestión energética de sistemas HVAC e iluminación en el ámbito educativo.

1 Objetivos

1.1 Objetivo General

Desarrollar un algoritmo basado en técnicas supervisadas para la detección y estimación de la ocupación, utilizando mediciones *in situ* de dos aulas de clase del edificio de la Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones (E3T).

1.2 Objetivos Específicos

Seleccionar las técnicas supervisadas más relevantes para la detección y estimación de la ocupación en edificaciones institucionales.

Definir métricas de evaluación que permitan medir y comparar el rendimiento de los modelos de detección de la ocupación.

Diseñar e implementar modelos de ocupación basados en técnicas supervisadas aplicados a casos de estudio en aulas de clase del edificio de la E3T.

Comparar el desempeño de los modelos implementados utilizando las métricas preestablecidas y seleccionar el modelo con mejor desempeño en términos de precisión y eficiencia.

2 Marco teórico

Este capítulo describe la relevancia de la eficiencia energética en edificaciones institucionales, los sensores y variables que permiten inferir la ocupación, así como las técnicas supervisadas empleadas en la literatura para la detección y la estimación de la ocupación. Finalmente, se presentan las técnicas de evaluación para evaluar el desempeño.

2.1 Eficiencia energética en edificaciones institucionales

A nivel global, el sector de las edificaciones representa aproximadamente el 30 % del consumo energético final y el 37 % de las emisiones de CO_2 relacionadas con la energía (International Energy Agency, 2023; United Nations Environment Programme, 2023). En entornos institucionales, la demanda se concentra principalmente en los sistemas de climatización y ventilación (40–60 %) e iluminación (15–20 %) (ASHRAE, 2021; U.S. Department of Energy, 2022). Este panorama ha impulsado compromisos internacionales como el Acuerdo de París (United Nations Framework Convention on Climate Change, 2015) y, en el contexto colombiano, un marco normativo robusto liderado por la Ley 1715 de 2014 y el reciente Decreto 2407 de 2024. Estas normativas fomentan la digitalización y la gestión inteligente de la demanda mediante sistemas de monitoreo y control para reducir la huella de carbono en espacios construidos (Decreto 2407, 2024; Ley 1715, 2014).

En este escenario, la identificación de los patrones reales de uso se posiciona como una estrategia crítica para superar la ineficiencia de la operación basada en horarios fijos. La integración de sistemas inteligentes apoyados en datos de presencia puede generar ahorros energéticos de entre el 10 % y el 30 % en edificios educativos (International Energy Agency, 2023), mientras que la adaptación dinámica de sistemas HVAC puede reducir el consumo anual entre un 15 % y un 40 % (ASHRAE, 2021). Por lo tanto, la detección y estimación de ocupación mediante sensores ambientales no intrusivos constituye un elemento fundamental para optimizar el confort térmico y la eficiencia operativa en infraestructuras existentes, permitiendo un ajuste dinámico de los recursos en función de la demanda real (Mena et al., 2022).

2.2 Sensores y variables para inferencia de ocupación

La literatura clasifica la gestión de ocupación en enfoques intrusivos y no intrusivos. Los métodos **intrusivos** (cámaras, biometría) ofrecen alta precisión, pero su viabilidad en entornos institucionales es limitada debido a costos, demanda computacional y riesgos críticos de privacidad ante oclusiones o cambios espaciales (Mena et al., 2022).

En contraste, los métodos **no intrusivos** infieren la ocupación mediante variables ambientales (CO_2, temperatura, humedad, movimiento) y eléctricas (potencia activa), facilitando la integración en infraestructuras existentes sin comprometer datos sensibles (Candanedo & Feldheim, 2016; Gunay et al., 2019). La fusión de estos sensores optimiza la detección al capturar patrones de exhalación y calor metabólico, superando el desempeño de modelos univariables (Chitnis et al., 2025; Mena et al., 2022; Mena-Martinez et al., 2024).

Este trabajo adopta un enfoque exclusivamente no intrusivo por restricciones de privacidad y disponibilidad técnica en el campus. Esta estrategia garantiza un balance óptimo entre precisión (detección binaria y por rangos) y protección de la información, aprovechando la correlación directa entre el comportamiento del espacio y las señales medidas.

2.3 Técnicas para la detección y estimación de ocupación

El aprendizaje supervisado modela la relación entre variables ambientales y el estado de ocupación a partir de datos históricos etiquetados (Mena et al., 2022; Rashid et al., 2018). Si bien su desempeño depende de la calidad y representatividad de los datos disponibles, el uso de algoritmos con distintos niveles de complejidad permite capturar relaciones de diversa naturaleza, ajustándose a los requerimientos de cada escenario (Candanedo & Feldheim, 2016; Gunay et al., 2019).

Regresión logística: Es un modelo de referencia para la clasificación binaria en detección de ocupación debido a su interpretabilidad y bajo costo computacional (Gunay et al., 2019; Hosmer et al., 2013). Estima la probabilidad condicional $P(y = r \mid \mathbf{x})$, con $r \in 0, 1$, mediante una función sigmoidal:

$$P(y = r \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}, \tag{1}$$

donde $\mathbf{w}$ y b representan los parámetros aprendidos (Hastie et al., 2017). No obstante, su capacidad para modelar relaciones no lineales complejas es limitada, lo que motiva el uso de modelos más expresivos (Candanedo & Feldheim, 2016; Rashid et al., 2018). La Figura A.1 (Anexo 5) ilustra la función sigmoide empleada.

K-Nearest Neighbors (KNN): Es un método no paramétrico basado en instancias que clasifica nuevas observaciones según su similitud con los datos de entrenamiento (Cover & Hart, 1967; Gunay et al., 2019). Dado un vector $\mathbf{x}_{\text{new}}$, el algoritmo identifica los k vecinos más cercanos mediante una métrica de distancia y asigna la clase predominante:

$$\hat{y} = \arg, \max c \sum i \in \mathcal{N}_k \mathbb{I}(y_i = c), \tag{2}$$

donde $\mathbb{I}(\cdot)$ es la función indicadora y k controla el compromiso entre sensibilidad al ruido y capacidad de generalización (Hastie et al., 2017). Si bien es intuitivo y fácil de implementar, su costo computacional aumenta con el tamaño del conjunto de datos y es sensible a la escala de las variables, por lo que requiere una normalización previa (Candanedo & Feldheim, 2016). La Figura A.2 (Anexo 5) ilustra su esquema conceptual.

Naïve Bayes: Es un clasificador probabilístico basado en el teorema de Bayes, ampliamente utilizado como modelo de referencia por su simplicidad y eficiencia computacional (Gunay et al., 2019; Mitchell, 1997). Bajo el supuesto de independencia condicional entre las d variables de entrada $\mathbf{x}$, estima la probabilidad del estado de ocupación y como:

$$P(y \mid \mathbf{x}) \propto P(y) \prod_{j=1}^{d} P(x_j \mid y), \tag{3}$$

donde $P(y)$ es la probabilidad a priori y $P(x_j \mid y)$ la verosimilitud de cada característica, modelada comúnmente mediante distribuciones gaussianas para variables continuas (Candanedo & Feldheim, 2016). Si bien este supuesto simplifica el cálculo, puede resultar restrictivo debido a la

correlación inherente entre las variables ambientales en edificaciones (Zhang & Jain, 2019). La Figura A.3 (Anexo 5) ilustra su principio de funcionamiento.

Árboles de decisión: Modelan relaciones no lineales mediante la partición recursiva del espacio de características en regiones homogéneas, mediante reglas jerárquicas de decisión (Breiman et al., 1984; Quinlan, 1986). Cada división se selecciona maximizando la ganancia de información (IG), que cuantifica la reducción de la entropía H:

$$\text{IG} = H(\mathcal{D}) - \sum_{k} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} H(\mathcal{D}_k). \tag{4}$$

Aunque son altamente interpretables, presentan sensibilidad al ruido y riesgo de sobreajuste, lo que motiva el uso de métodos de ensamble, como Random Forest, para mejorar la robustez y la capacidad de generalización (Breiman, 2001; Li & Fan, 2020). La Figura A.4 (Anexo 5) muestra su estructura conceptual.

Máquinas de Soporte Vectorial (SVM): Destacan por su capacidad para manejar espacios de alta dimensionalidad y construir fronteras de decisión robustas mediante la maximización del margen entre clases (Cortes & Vapnik, 1995; Gunay et al., 2019). En su formulación lineal, el problema de optimización consiste en:

$$\min_{\mathbf{w},b} \ \frac{1}{2}|\mathbf{w}|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \tag{5}$$

donde $y_i \in -1, 1$ representa el estado de ocupación. Para capturar relaciones no lineales, las SVM emplean funciones *kernel*, siendo las más utilizadas las lineal, polinómica y de base radial (RBF) (Cortes & Vapnik, 1995; Gunay et al., 2019). Aunque ofrecen un alto poder de generalización, los kernels no lineales incrementan la sensibilidad a los hiperparámetros de regularización y a la escala de las variables, lo que requiere una validación cuidadosa (Chitnis et al., 2025; Gunay et al., 2019). La Figura A.5 (Anexo 5) ilustra el principio de funcionamiento de la SVM.

Random Forest (RF): Es un método de ensamble que combina múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de datos y variables, con el fin de mejorar la estabi-

lidad y la robustez del modelo (Breiman, 2001). En clasificación, la predicción final $\hat{y}$ se obtiene mediante votación mayoritaria entre los M árboles:

$$\hat{y} = \operatorname{mode} h_1(\mathbf{x}), \ldots, h_M(\mathbf{x}), \tag{6}$$

donde $h_m(\mathbf{x})$ es la salida del árbol m-ésimo. Este esquema reduce la varianza y mitiga el efecto del ruido ambiental, permitiendo capturar relaciones no lineales complejas en datos de sensores (Candanedo & Feldheim, 2016; Gunay et al., 2019). Aunque presenta mayor costo computacional y menor interpretabilidad que un árbol individual, su alta capacidad de generalización respalda su uso extendido en edificaciones inteligentes (Chitnis et al., 2025; Mena et al., 2022). La Figura A.6 (Anexo 5) ilustra su funcionamiento.

Multi-layer Perceptron (MLP): Es una red neuronal de alimentación directa que modela relaciones no lineales mediante una arquitectura de capas interconectadas (Haykin, 2009; Rumelhart et al., 1986). Cada capa l transforma las activaciones previas $\mathbf{h}^{(l-1)}$ como:

$$\mathbf{h}^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right), \tag{7}$$

donde $\mathbf{W}^{(l)}$ y $\mathbf{b}^{(l)}$ son los parámetros aprendidos y $\sigma(\cdot)$ es una función de activación no lineal (Haykin, 2009). Su flexibilidad permite capturar interacciones complejas entre variables ambientales; sin embargo, su desempeño depende fuertemente de la arquitectura y los hiperparámetros, y su interpretabilidad es limitada (J. Chen et al., 2018). Aun así, su capacidad de generalización justifica su uso en aplicaciones avanzadas de edificios inteligentes (Mena et al., 2022). La Figura A.7 (Anexo 5) esquematiza su arquitectura.

2.3.1 Criterios de selección

Aunque el estado del arte destaca el uso de regresión logística, k-NN, Naïve Bayes y árboles de decisión por su interpretabilidad, estos métodos presentan limitaciones críticas ante relaciones no lineales y desbalance de clases (Candanedo & Feldheim, 2016; Gunay et al., 2019). Por ello, se seleccionaron SVM, RF y MLP para la fase experimental, dada su probada robustez y precisión

en escenarios ambientales complejos (Chitnis et al., 2025; Mena-Martinez et al., 2024), resultando idóneos para capturar la variabilidad de los patrones de ocupación. La comparación detallada de sus características se presenta en la Tabla A.1 (Anexo 5).

2.4 Métricas de evaluación

La evaluación del desempeño analiza la consistencia de los modelos frente al ruido y al desbalance característico de los datos de sensores en edificaciones (Chitnis et al., 2025; Mohammadabadi et al., 2022). Para ello, se emplean métricas de evaluación derivadas de la matriz de confusión, ampliamente utilizadas en problemas de detección y estimación de ocupación.

2.4.1 Métricas para clasificación binaria

En la detección binaria de ocupación (*ocupado/desocupado*), el desempeño se evalúa mediante **Accuracy**, **Precision**, **Recall**, **F1-score** y **AUC-ROC**. Estas métricas permiten analizar de forma complementaria la exactitud global, la confiabilidad de las predicciones positivas y la capacidad del modelo para detectar correctamente la ocupación, mitigando el impacto del desbalance de clases (Candanedo & Feldheim, 2016).

Matriz de confusión: Resume el desempeño del clasificador organizando las predicciones frente a las etiquetas reales. En clasificación binaria, se define como:

	Predicción = 1	Predicción = 0
Real = 1	TP	FN
Real = 0	FP	TN

donde TP, TN, FP y FN representan verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, respectivamente.

Exactitud (Accuracy): Mide la proporción total de predicciones correctas:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

Aunque es una métrica global intuitiva, puede resultar poco representativa en escenarios

con desbalance de clases.

Precisión (Precision) y Sensibilidad (Recall): La Precision evalúa la confiabilidad de las predicciones positivas:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{9}$$

mientras que el Recall cuantifica la capacidad del modelo para identificar correctamente los estados de ocupación:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{10}$$

Estas métricas son especialmente relevantes en aplicaciones de gestión energética, donde los falsos negativos pueden afectar el confort y la eficiencia del sistema HVAC.

Medida F1: Proporciona un compromiso entre precisión y sensibilidad mediante su media armónica:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{11}$$

siendo adecuada para conjuntos de datos desbalanceados.

Curva ROC y AUC: La curva ROC relaciona la tasa de verdaderos positivos ($TPR =$ Recall) con la tasa de falsos positivos ($FPR = \frac{FP}{FP+TN}$) bajo distintos umbrales de decisión. El Área Bajo la Curva (**AUC**) resume la capacidad discriminativa del clasificador, donde valores cercanos a 1 indican alta separabilidad entre clases (Mohammadabadi et al., 2022).

2.4.2 Métricas para clasificación multiclase

Para la estimación de múltiples niveles de ocupación, se emplean métricas que capturan tanto el rendimiento global como el comportamiento frente a clases desbalanceadas. En este contexto, *Precision*, *Recall* y *F1-score* se calculan por clase y se agregan mediante esquemas de promediado.

- **Accuracy global**, definida como la proporción total de predicciones correctas:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i), \tag{12}$$

donde N es el número total de muestras.

- **Precision, Recall y F1-score ponderados**, agregados según el soporte de cada clase:

$$\text{F1}_{\text{weighted}} = \sum_{k=1}^{K} w_k \cdot \text{F1}_k, \tag{13}$$

con $w_k = \frac{n_k}{N}$.

- **Área bajo la curva ROC (AUC) multiclase**, calculada mediante la estrategia *One-vs-Rest* (OvR). Se reporta el promedio macro (*Macro-average AUC*), el cual calcula la media aritmética de los AUCs de todas las clases:

$$\text{AUC}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} \text{AUC}(\text{Clase}_k \text{ vs Resto}). \tag{14}$$

Se seleccionó este promedio en lugar del ponderado para evaluar el desempeño equitativamente, evitando que la clase mayoritaria (R0) enmascare el rendimiento en las clases críticas.

- **Matriz de confusión multiclase**, que permite identificar errores sistemáticos entre niveles de ocupación adyacentes (Mena-Martinez et al., 2024).

2.4.3 Criterios de selección

La selección de métricas depende del problema, las características de los datos y los requisitos operativos. En edificaciones, deben considerarse el desbalance de clases, la sensibilidad para el control HVAC, el ruido de los sensores y la interpretabilidad en sistemas de gestión energética.

En este trabajo, para la detección binaria, se priorizan **Accuracy, Precision, Recall y F1-score** para mitigar el impacto del desbalance y los falsos negativos en la eficiencia energética. En la clasificación multiclase, se emplean el **F1-score ponderado** y la **matriz de confusión**, permitiendo evaluar el desempeño frente a niveles de ocupación adyacentes y garantizar una operación coherente del edificio inteligente.

3 Metodología

Este capítulo describe la metodología para el desarrollo y evaluación de los modelos de detección y estimación de ocupación. El enfoque se basa en el análisis de datos recolectados mediante sensores ambientales y energéticos. Se detallan las etapas de preprocesamiento, selección de variables, definición de configuraciones y criterios de evaluación. Asimismo, se presentan las métricas utilizadas para sustentar objetivamente las decisiones de diseño adoptadas en el trabajo.

3.1 Caso de estudio

El caso de estudio se desarrolló en dos aulas ubicadas en el cuarto piso de la Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones (E3T) de la Universidad Industrial de Santander, espacios utilizados de forma regular para actividades académicas. Esta condición permitió capturar variaciones representativas en los patrones de ocupación a lo largo del semestre.

El monitoreo se llevó a cabo entre el 16 de febrero y el 30 de abril de 2024, mediante sensores IoT instalados de forma permanente en las aulas, con un intervalo de muestreo de 5 minutos, lo que dio lugar a 21 382 observaciones por variable. El periodo analizado incluyó semanas con actividad académica regular, evaluaciones parciales y días sin programación, lo que aportó una variabilidad temporal adecuada para el análisis de la ocupación. Las figuras B.1 y B.2 (Anexo 5) presentan el aula 404 y el esquema de monitoreo, respectivamente, tomados de estudios previos realizados en el mismo edificio (Ortega-Diaz et al., 2025).

Las variables empleadas corresponden a mediciones ambientales y eléctricas no intrusivas asociadas al uso del espacio. Del conjunto total disponible se seleccionaron aquellas con mayor relación directa con la presencia de personas: potencia activa, estado de puertas y ventanas, temperatura interior, humedad interior y detección de movimiento. La descripción completa de las variables registradas se presenta en la Tabla B.1 (Anexo 5).

A partir del conjunto total de variables disponibles, se seleccionó un subconjunto de características directamente relacionadas con la detección y estimación de la ocupación. Esta selección se fundamentó en su relación física y operativa con la presencia de ocupantes en el aula, así como

en la existencia de dependencia y redundancia entre algunas variables originalmente disponibles. En consecuencia, se priorizaron aquellas variables que aportan información relevante y complementaria, evitando la inclusión de características altamente correlacionadas entre sí.

Bajo estas consideraciones, en la Tabla B.2 (Anexo 5) se presentan las variables usadas.

Para el entrenamiento y evaluación de los modelos supervisados se definió una referencia del estado real de ocupación (*ground truth*). Las etiquetas de ocupación se obtuvieron mediante un proceso de anotación basado en los registros de las cámaras instaladas en cada aula, realizando, para cada intervalo de muestreo, el conteo de personas y computadores en uso, en correspondencia temporal con las mediciones de los sensores.

Con el fin de ilustrar el comportamiento típico de las variables seleccionadas, se analizó una semana representativa del periodo de monitoreo (del 19 al 26 de febrero de 2024). La Figura B.3 (Anexo 5) muestra la evolución temporal de estas variables y permite identificar patrones asociados a los ciclos diarios de ocupación y su relación con las condiciones ambientales y eléctricas.

3.2 Preprocesamiento de los datos

El preprocesamiento inició con la limpieza de registros, orientada a detectar y eliminar valores faltantes, duplicados o inconsistentes, a fin de garantizar la integridad del conjunto de datos y evitar sesgos durante el entrenamiento de los modelos. Adicionalmente, las variables categóricas correspondientes al estado de puertas y ventanas fueron codificadas mediante esquemas numéricos, lo que asegura su compatibilidad con los algoritmos de aprendizaje supervisado seleccionados.

Para la división del conjunto de datos se empleó un esquema de partición aleatoria estratificada (*Stratified Shuffle Split*). Esta decisión metodológica se priorizó sobre la división puramente temporal debido al severo desbalance de clases (donde rangos como R5 representan $< 1\,\%$ de la muestra). La estratificación garantiza que el conjunto de prueba ($30\,\%$) contenga una representación proporcional de todos los niveles de ocupación, evitando que los eventos menos frecuentes queden excluidos de la evaluación. La capacidad de generalización temporal y espacial del modelo se valida posteriormente mediante su aplicación en un escenario totalmente independiente (Aula 405).

Posteriormente, las variables continuas fueron normalizadas mediante el método *StandardScaler*, transformando los datos a media cero y desviación estándar unitaria. Este procedimiento resulta particularmente relevante para modelos sensibles a la escala de las variables, como las SVM y el MLP, lo que permite un tratamiento homogéneo de magnitudes medidas por sensores de naturalezas diferentes.

Dado que el objetivo del estudio es evaluar el desempeño de los modelos en condiciones representativas de operación real, no se aplicaron técnicas de remuestreo para abordar el desbalance de clases. En su lugar, este aspecto se consideró en la etapa de evaluación mediante métricas como Precision, Recall y F1-score. Finalmente, se consolidó un conjunto de datos preprocesado con las variables seleccionadas, garantizando coherencia temporal y una representación fiel de las condiciones reales del entorno analizado. El procedimiento seguido en esta etapa se resume en la Tabla B.3.

3.3 Selección de modelos y configuración de hiperparámetros

En coherencia con el marco teórico y los objetivos del estudio, se seleccionaron tres modelos de aprendizaje supervisado: Máquinas de Soporte Vectorial (SVM), Random Forest (RF) y Multi-layer Perceptron (MLP). La elección se fundamenta en su capacidad probada para modelar relaciones no lineales en edificaciones institucionales mediante sensores no intrusivos. Estos modelos ofrecen la flexibilidad necesaria para aplicarse tanto a la detección como a la estimación de la ocupación, lo que permite una comparación metodológica homogénea. Su estructura permite abordar la variabilidad temporal de los patrones de ocupación y las condiciones ambientales, consolidándose como los modelos base para el proceso experimental detallado en las siguientes secciones.

El desempeño de los modelos supervisados se optimizó mediante la evaluación de hiperparámetros definidos según rangos reportados en la literatura para sensores ambientales y eléctricos. Para SVM, se analizaron kernels lineal, polinomial y RBF con valores de $C \in \{1, 10\}$, buscando equilibrar la regularización y el riesgo de sobreajuste (Candanedo & Feldheim, 2016; Chaudhari et al., 2024; Rashid et al., 2018).

En el caso de RF, se evaluaron configuraciones de 100, 200 y 500 árboles (*n_estimators*) para reducir la varianza y estabilizar el error sin exceder el costo computacional (Breiman, 2001; Gunay et al., 2019; Li & Fan, 2020). Para el MLP, se emplearon arquitecturas de dos capas ocultas con un número moderado de neuronas e iteraciones máximas variables, garantizando la convergencia y captura de relaciones no lineales sin sobreajuste (J. Chen et al., 2018; Mohammadabadi et al., 2022). Todas las configuraciones se evaluaron bajo un esquema experimental uniforme para asegurar la comparabilidad y consistencia de los resultados.

Los resultados de estas configuraciones se analizan en el **Capítulo 4**, identificando el equilibrio entre desempeño y eficiencia. Para la validación, el dataset se dividió en entrenamiento (70 %) y evaluación (30 %), empleando observaciones no vistas para estimar la capacidad de generalización del modelo con un bajo costo computacional.

3.4 Detección de la ocupación

Esta tarea se define como una clasificación binaria donde la variable $y \in \{0, 1\}$ indica ocupación (1) o desocupación (0). El modelo utiliza un vector $\mathbf{x} = [T, H, P, D, W, M]^\top$ con variables de temperatura, humedad, potencia, puerta, ventanas y movimiento. El objetivo es aprender una función $f : \mathbb{R}^6 \to \{0, 1\}$ que genere una predicción $\hat{y} = f(\mathbf{x})$ mediante la regla:

$$\hat{y} = \begin{cases} 1 & \text{si } \mathbb{P}(y = 1 \mid \mathbf{x}) \geq 0.5 \\ 0 & \text{en otro caso} \end{cases} \tag{15}$$

Este proceso de inferencia se detalla en la Tabla 1.

Tabla 1

Proceso de estimación del nivel de ocupación por rangos.

Paso	Descripción
1	Codificar variables categóricas
2	Normalizar variables continuas (μ, σ)
3	Ejecutar modelo multiclase f
4	Calcular probabilidades $\mathbf{p} = f(\mathbf{x})$
5	Asignar rango $\hat{y} = \arg\max \mathbf{p}$

Esquema general del proceso de detección: El procedimiento completo para la detección de ocupación se resume mediante el pseudocódigo, el cual describe las etapas principales desde el preprocesamiento de los datos hasta la obtención de la predicción final del estado del aula (**Algoritmo 2**).

Este esquema constituye la base del sistema de detección basado en datos, ya que permite determinar de forma automatizada el estado de ocupación del aula a partir de mediciones ambientales y eléctricas no intrusivas.

El desempeño se evaluó mediante métricas de clasificación binaria: Accuracy, AUC, Recall y F1-score. Estas permiten analizar el rendimiento global y el impacto de errores específicos, como los falsos negativos, críticos en gestión energética y confort térmico.

Para la validación, se empleó una partición fija del 70 % de los datos para entrenamiento y 30 % para prueba con datos no vistos. Este esquema garantiza una evaluación consistente y de bajo costo computacional, facilitando la comparación directa entre los modelos supervisados aplicados al entorno universitario.

Algoritmo 2: Proceso de inferencia para detección de ocupación

Entrada: Vector de características $\mathbf{x}$
Salida: Predicción de ocupación $\hat{y}$
1. Cargar parámetros de normalización (μ, σ)
2. Normalizar entrada: $\mathbf{x} \leftarrow (\mathbf{x} - \mu)/\sigma$
3. Calcular probabilidad de ocupación $p \leftarrow f(\mathbf{x})$
4. **Si** $p \geq 0.5$ **entonces**:
   $\hat{y} \leftarrow 1$ // *Ocupado*
5. **Sino**:
   $\hat{y} \leftarrow 0$ // *Desocupado*
6. **Retornar** $\hat{y}$

3.4.1 Desbalance de clases en la detección de la ocupación

Para la tarea de detección de la ocupación se formuló un problema binario, donde la clase 0 corresponde a la ausencia de ocupación (0 personas) y la clase 1 agrupa todos los casos con presencia de ocupación (una o más personas).

En ambos conjuntos de datos se observa un desbalance significativo entre las clases. En el aula 404, la clase 0 representa aproximadamente el 80 % de las muestras, mientras que la clase 1 corresponde al 20 % restante. De manera similar, en el aula 405 la clase 0 concentra cerca del

81 % de los datos, y la clase 1 alrededor del 19 %.

Este comportamiento está asociado al uso real de los espacios académicos, donde los periodos sin ocupación son predominantes. En consecuencia, no se aplicaron técnicas de balanceo o remuestreo, con el objetivo de evaluar el desempeño de los modelos bajo condiciones representativas de operación real.

La Tabla C.1 presenta la distribución porcentual de las clases para la detección de la ocupación en las aulas 404 y 405. En ambos casos se observa un desbalance marcado entre la ausencia y la presencia de ocupación, coherente con el uso real de los espacios académicos.

3.5 Estimación de la ocupación

La segunda tarea consiste en la estimación del número de ocupantes mediante **clasificación multiclase**, donde $y \in \{0, 1, \ldots, 5\}$ representa seis niveles de ocupación: **R0**: 0 pers., **R1**: 1–7, **R2**: 8–14, **R3**: 15–21, **R4**: 22–28 y **R5**: 29–35 personas.

El modelo busca aprender una función $f : \mathbb{R}^n \longrightarrow \{0, \ldots, 5\}$ tal que $\hat{y} = f(\mathbf{x})$ determine el nivel de ocupación a partir de las variables ambientales y eléctricas.

La evaluación analizó la capacidad de los modelos para discriminar el estado de ocupación ($y \in \{0, 1\}$) y estimar sus niveles mediante métricas de clasificación y gráficos de *radar*. A partir de las predicciones en el conjunto de prueba, se construyó la matriz de confusión para identificar aciertos y errores, análisis crítico debido al desbalance natural de clases en entornos académicos.

Se calcularon las métricas Accuracy, AUC, Specificity, Recall y F1-score, permitiendo evaluar el rendimiento global y la minimización de falsos negativos. Asimismo, se integró el análisis de la curva ROC y el área bajo la curva (AUC) para contrastar la capacidad discriminativa bajo distintos umbrales de decisión. Los resultados y su comparativa técnica se detallan en el **Capítulo 4**.

3.5.1 Desbalance de clases por rangos de ocupación

En el aula 404, el rango R0 concentra aproximadamente el 80 % de las muestras, mientras que los rangos asociados a ocupaciones elevadas, en particular R5, representan menos del 1 % del

total. De forma similar, en el aula 405 el rango R0 agrupa más del 81 % de los datos, evidenciando una baja representación de los rangos superiores.

Este desbalance refleja condiciones reales de uso de los espacios académicos. Por esta razón, y con el fin de evaluar el desempeño de los modelos bajo escenarios reales de operación, no se aplicaron técnicas de remuestreo ni balanceo de clases.

La Tabla C.2 resume la distribución porcentual de las clases definidas por rangos de ocupación para la tarea de estimación de la ocupación en las aulas 404 y 405. Se evidencia un desbalance progresivo, donde los rangos de menor ocupación concentran la mayor proporción de muestras.

3.6 Validación comparativa con técnicas de Boosting

Con el objetivo de validar la robustez del modelo, que se va a seleccionar a continuación, frente al estado del arte en clasificación de datos tabulares, se incluyó un experimento comparativo utilizando *Extreme Gradient Boosting* (XGBoost). Este algoritmo, perteneciente a la familia de métodos de *boosting*, construye un ensamble secuencial donde cada nuevo árbol corrige los errores residuales de los anteriores, lo que suele ofrecer ventajas en datasets con fronteras de decisión complejas (T. Chen & Guestrin, 2016).

Durante la revisión del estado del arte y la selección de algoritmos supervisados, no se incluyeron inicialmente técnicas de *boosting* como XGBoost, debido a que el objetivo del diseño era comparar el desempeño de modelos representativos de distintas familias de aprendizaje supervisado, en lugar de evaluar múltiples variantes dentro de una misma familia. En este contexto, *RF* fue seleccionado como representante de los métodos de ensamble basados en árboles, dado su comportamiento robusto frente al ruido, su menor sensibilidad a la calibración de hiperparámetros y su desempeño consistente en problemas con desbalance de clases.

Adicionalmente, el alcance del proyecto se centró en el análisis comparativo entre enfoques lineales, basados en distancia, redes neuronales y ensambles de árboles, con el fin de identificar una arquitectura adecuada para la detección y estimación de ocupación en entornos instrumentados. Bajo esta lógica, XGBoost no fue considerado en la fase inicial, dado que su inclusión no aportaba diversidad metodológica adicional respecto a RF.

4 Resultados y análisis

Esta sección analiza los resultados obtenidos por los algoritmos de aprendizaje supervisado evaluados para la detección y estimación de ocupación. La comparación se realiza a partir de métricas de desempeño que permiten evaluar tanto la capacidad de clasificación como el equilibrio entre aciertos y errores.

4.1 Resultados para el aula 404

4.1.1 Análisis de desempeño de detección de ocupación

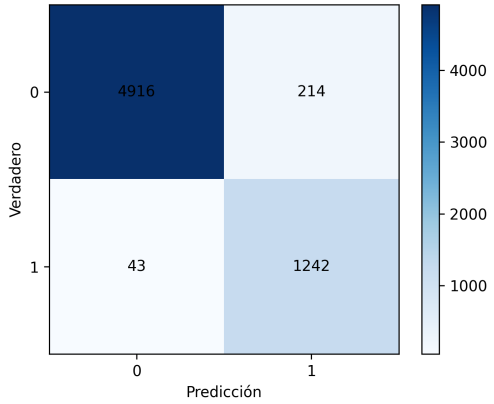
- SVM

La evaluación de las configuraciones SVM, detallada en la Tabla D.1, demuestra que la complejidad de la frontera de decisión es determinante para la detección de ocupación. Los kernels lineales presentan un desempeño insuficiente (promedio ≈ 0.937), evidenciando que las relaciones entre variables ambientales no son linealmente separables. Si bien los kernels polinomiales mejoran la capacidad de modelado (≈ 0.949), implican un costo computacional elevado sin alcanzar el óptimo. En contraste, como se ilustra en la comparativa de la Figura F.1 (Anexo 5), el kernel RBF con $C = 10$ maximiza el rendimiento global (0.954), logrando el mejor equilibrio entre *Accuracy*, *Specificity* y *Recall*.

La validación mediante la matriz de confusión (Figura 1) ratifica la robustez de la configuración RBF ($C = 10$). La alta concentración de aciertos en la diagonal principal confirma su capacidad discriminativa, mientras que el análisis de los errores revela que estos se limitan casi exclusivamente a las observaciones cercanas al umbral de decisión.

Figura 1

Matriz de confusión del modelo SVM para la detección de ocupación (aula 404).



Este comportamiento es técnicamente consistente con los periodos de transición física (entrada y salida de ocupantes), permitiendo concluir que esta configuración ofrece el balance idóneo entre precisión y generalización para el entorno del aula 404.

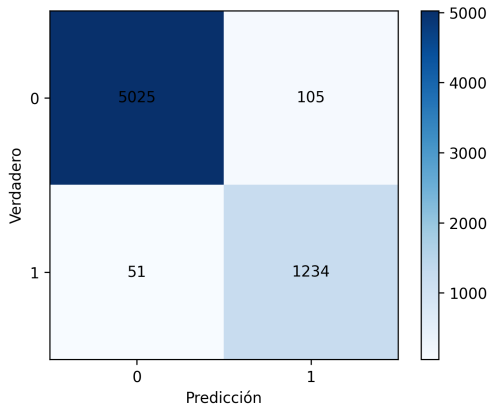
- RF

El análisis del modelo RF (Tabla D.2) evidencia una alta capacidad de discriminación, destacándose por la consistencia en todas las métricas evaluadas. Respecto a la optimización de hiperparámetros, los datos revelan un punto de saturación en el aprendizaje: el desempeño mejora progresivamente hasta alcanzar un máximo promedio de 0.970 con 200 árboles. Como se ilustra en la tendencia asintótica de la Figura F.2 (Anexo 5), incrementar el número de estimadores más allá de este umbral eleva el costo computacional sin aportar beneficios significativos a la clasificación.

La robustez de esta configuración se ratifica mediante la matriz de confusión (Figura 2). La alta densidad de aciertos en la diagonal principal confirma una detección precisa tanto de eventos de ocupación como de vacancia, mientras que los errores residuales se limitan exclusivamente a las zonas de transición temporal propias de la dinámica de uso del aula. En conclusión, la arquitectura de 200 árboles constituye el compromiso óptimo entre robustez predictiva y eficiencia algorítmica para el escenario evaluado.

Figura 2

Matriz de confusión del modelo RF para la detección de ocupación (aula 404).



- MLP

Los resultados del modelo MLP para el aula 404 (Tabla D.3) demuestran que las arquitecturas evaluadas mantienen un desempeño elevado y consistente. Esto confirma la alta capacidad del sistema para diferenciar con precisión entre los estados de ocupación y no ocupación. La arquitectura (50, 25) destaca por alcanzar los valores más altos en *Accuracy* (0.969), *Specificity* (0.978), *F1-score* (0.922) y AUC (0.99), registrando además el menor tiempo de ejecución. Este rendimiento demuestra un equilibrio óptimo entre eficacia de clasificación y eficiencia computacional.

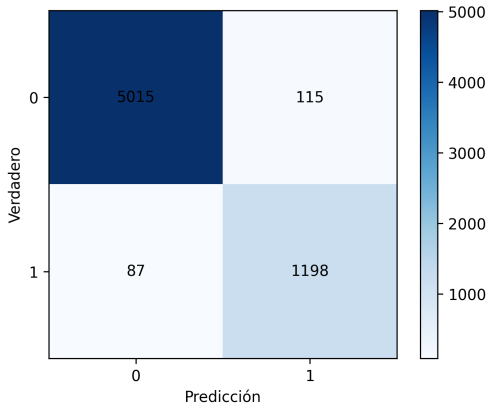
Aunque configuraciones más profundas alcanzan métricas similares, no ofrecen mejoras sustanciales y sí elevan los tiempos de entrenamiento. Esto demuestra que, para los datos analizados, el incremento en la complejidad del modelo no aporta ventajas significativas frente a arquitecturas más sencillas. Para evaluar globalmente cada arquitectura, se analizaron los promedios de las métricas detallados en la Figura F.3 (Anexo 5). El análisis del impacto de la profundidad y densidad de la red confirmó que la configuración (50, 25) obtiene el mejor rendimiento promedio, consolidándose como la arquitectura de referencia para la detección de ocupación en el aula 404.

La matriz de confusión del modelo MLP (50, 25), detallada en la Figura 3, refleja un alto índice de aciertos en ambas clases al concentrar las muestras en la diagonal principal. Los errores marginales se limitan principalmente a las transiciones cercanas al umbral de decisión entre los

estados de ocupación.

Figura 3

Matriz de confusión del modelo MLP para la detección de ocupación (aula 404).



4.1.2 Análisis de desempeño final utilizando Radar Chart Aula 404

Con el propósito de sintetizar el comportamiento de los modelos en la detección de ocupación, se empleó un gráfico de tipo *radar* como herramienta de comparación visual integrada.La representación mediante el gráfico de radar (Figura F.7, Anexo 5) permite contrastar simultáneamente las métricas de *Accuracy*, *Recall*, *Specificity*, *F1-score* y *AUC*, facilitando la identificación de sesgos o fortalezas globales en los enfoques analizados.

En términos generales, los tres modelos exhiben un desempeño sobresaliente y consistente, con valores elevados que reflejan una alta fiabilidad en la discriminación de los estados de ocupación y vacancia. No obstante, el modelo RF presenta la cobertura de área más amplia y uniforme, lo que técnicamente se traduce en un equilibrio superior entre la sensibilidad y la especificidad. Por su parte, SVM y MLP muestran perfiles altamente competitivos, aunque con ligeras variaciones en el *Recall* y el *F1-score*, sugiriendo una sensibilidad marginalmente menor en la identificación del estado ocupado frente a RF.

En síntesis, la comparativa visual mediante el radar chart ratifica los hallazgos cuantitativos y consolida la selección de los modelos evaluados como herramientas robustas para el entorno del aula 404. Esta herramienta gráfica permite concluir que, si bien las diferencias numéricas son

sutiles, la estabilidad métrica de RF lo posiciona como la arquitectura de referencia para la implementación del algoritmo de detección.

4.1.3 Análisis de desempeño para estimación de ocupación

- SVM

La Tabla D.4 resume los resultados de las SVM para la estimación multiclase del nivel de ocupación en el aula 404. Las métricas reportadas representan los valores promedio obtenidos mediante el esquema de evaluación definido previamente. Las configuraciones con núcleo lineal presentan los valores más bajos en *Accuracy*, *Recall* y *F1-score*, evidenciando una capacidad limitada para modelar las relaciones no lineales de los datos. En contraste, el uso de núcleos no lineales permite una mejora progresiva en el desempeño, destacando especialmente los resultados obtenidos con el núcleo RBF. La configuración SVM con núcleo RBF y $C = 10$ logra el mejor desempeño del conjunto con un *Accuracy* de 0.49 y un AUC de 0.758. Estos valores indican una mayor capacidad discriminativa, aunque este incremento en la eficacia conlleva un costo computacional superior en comparación con los otros algoritmos evaluados.

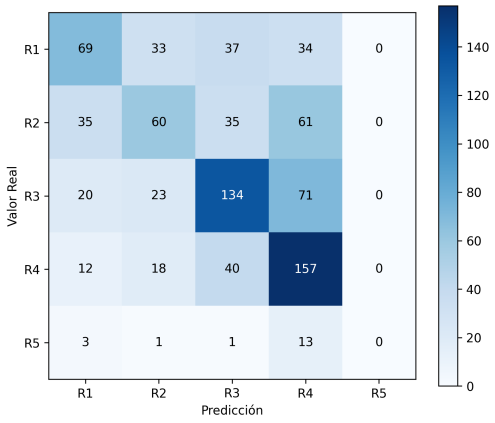
Para facilitar la comparación global entre configuraciones, se empleó el valor promedio de las métricas de desempeño, criterio que permite evaluar conjuntamente el efecto del tipo de núcleo y del parámetro C. Como se detalla en la Figura F.4, los resultados confirman que las configuraciones con núcleos no lineales, especialmente el RBF con $C = 10$, presentan el mejor balance general para la estimación del nivel de ocupación en el aula 404.

La matriz de confusión del modelo SVM (RBF, $C = 10$), detallada en la Figura 4, muestra un mejor desempeño en las clases intermedias R3 y R4, donde se concentra el mayor número de aciertos. Este comportamiento sugiere que el modelo captura con mayor efectividad los patrones de ocupación media, los cuales poseen una mayor representación en el conjunto de datos. En contraste, las clases correspondientes a niveles bajos y muy altos de ocupación (R1, R2 y R5) presentan mayores confusiones, principalmente hacia clases adyacentes. Este comportamiento se asocia al carácter gradual del nivel de ocupación y al desbalance de datos presente en los rangos extremos. En general, los errores observados se concentran entre clases contiguas, lo cual resulta

coherente con la naturaleza ordinal del problema de estimación del nivel de ocupación.

Figura 4

Matriz de confusión del modelo SVM para la estimación de ocupación (aula 404).



- RF

La Tabla D.5 presenta los resultados obtenidos con el algoritmo RF para distintos números de árboles en la tarea de estimación del nivel de ocupación . La comparación entre configuraciones se realizó a partir de las métricas promedio de desempeño, considerando el enfoque de clasificación multiclase definido previamente. En términos generales, RF presenta un desempeño consistente y superior al observado con las SVM para esta tarea, particularmente en las métricas de *Recall* y *F1-score*. Este comportamiento indica una mayor capacidad del modelo para discriminar entre los distintos rangos de ocupación, logrando una mejor identificación de los niveles con mayor representación en el conjunto de datos del aula 404.

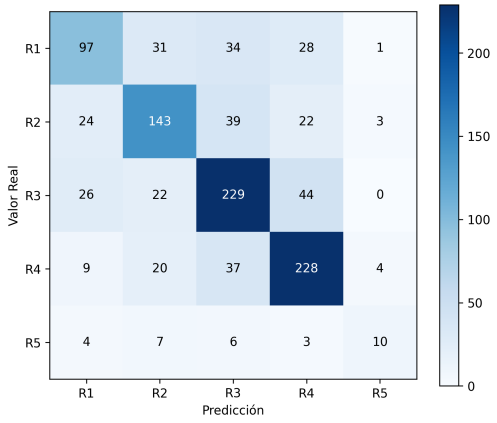
La Figura F.5 presenta la comparación gráfica del desempeño promedio para las distintas configuraciones de RF. En esta representación se evidencia que el modelo alcanza su mejor comportamiento con 200 árboles, mientras que un mayor número de estos no aporta mejoras relevantes y tiende a estabilizar el desempeño, confirmando que una complejidad moderada es suficiente para capturar los patrones de los datos. La configuración con 200 árboles alcanza un *Accuracy* de 0.66, un *F1-score* de 0.604 y un AUC de 0.882, manteniendo además un tiempo de ejecución reducido.

La matriz de confusión del modelo RF con 200 árboles, presentada en la Figura 5, evidencia

un buen desempeño general en la estimación del nivel de ocupación, con una clara concentración de aciertos sobre la diagonal principal. En particular, las clases intermedias (R2, R3 y R4) presentan los mayores porcentajes de clasificación correcta, lo que indica que el modelo logra capturar adecuadamente los patrones asociados a los niveles de ocupación más frecuentes en el conjunto de datos analizado.

Figura 5

Matriz de confusión del modelo RF para la estimación de ocupación (aula 404).



Los errores de clasificación se concentran principalmente entre clases adyacentes, como R2–R3 y R3–R4, lo cual es coherente con la naturaleza gradual de la ocupación. En contraste, las confusiones hacia clases extremas son menos frecuentes, sugiriendo una discriminación global adecuada. En conjunto, estos resultados posicionan a RF como una alternativa robusta para la estimación del nivel de ocupación, especialmente en los rangos medios predominantes del aula 404.

- MLP

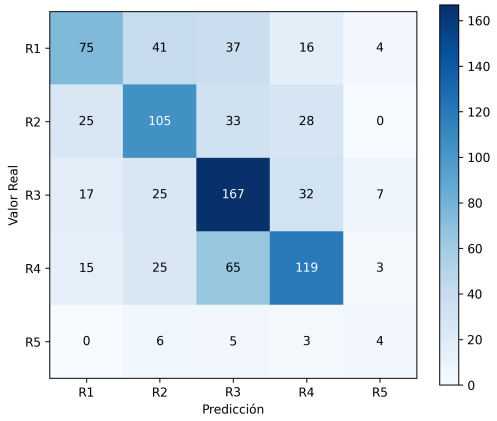
El análisis de los modelos MLP (Tabla D.6) revela que la capacidad de generalización del algoritmo se ve favorecida por el incremento en la profundidad de la red, siendo la arquitectura (200, 200, 50) la configuración óptima con un *Accuracy* de 0.49 y un *F1-score* de 0.384. No obstante, al contrastar estos resultados con los obtenidos por RF, se evidencia que MLP presenta un desempeño

global inferior y una mayor demanda de recursos computacionales, tendencia que se ratifica en la comparativa de promedios presentada en la Figura F.6 (Anexo 5).

Argumentalmente, la matriz de confusión (Figura 6) permite identificar una segmentación clara en el desempeño: mientras que las clases intermedias (R3 y R4) muestran una precisión aceptable, los niveles de ocupación bajos (R1 y R2) presentan una alta tasa de solapamiento con sus clases adyacentes. Esta limitación técnica, sumada a la nulidad predictiva en el nivel máximo (R5), se atribuye directamente al desbalance en la distribución del dataset y a la cercanía semántica entre rangos de ocupación.

Figura 6

Matriz de confusión del modelo MLP para la estimación de ocupación (aula 404).



En consecuencia, aunque MLP logra capturar la naturaleza gradual del fenómeno, su sensibilidad ante la variabilidad de los datos en los extremos del espectro reduce su fiabilidad como clasificador multiclase frente a modelos de ensamble.

4.1.4 Análisis de desempeño final mediante Radar Chart - Aula 404

Para realizar una comparación visual integrada de la tarea de estimación en el aula 404, se empleó un gráfico de tipo *radar* construido a partir de métricas normalizadas. Este análisis permite contrastar las configuraciones óptimas identificadas: SVM (RBF, $C = 10$), RF (200 árboles) y MLP (200, 100, 50).

Como se detalla en la Figura F.8 (Anexo 5), esta representación gráfica permite visuali-

zar de forma simultánea el *Accuracy*, *Recall*, *F1-score* y *AUC*. En este escenario, el modelo RF sobresale por alcanzar valores más elevados y homogéneos, cubriendo una mayor área de superficie; dicho comportamiento es coherente con su superior capacidad de discriminación en rangos de ocupación intermedios. Por su parte, los modelos SVM y MLP exhiben áreas más reducidas y magnitudes comparables entre sí, lo que refleja un desempeño limitado en la estimación multiclase. Sus principales restricciones se asocian a una menor sensibilidad en clases específicas y a la dificultad técnica para diferenciar rangos de ocupación adyacentes bajo las condiciones evaluadas.

En conjunto, el análisis mediante el gráfico de radar complementa los resultados cuantitativos y facilita la comparación directa entre modelos. Esta comparativa permite identificar con claridad los compromisos entre el desempeño predictivo, la estabilidad de las métricas y la complejidad computacional para la estimación del nivel de ocupación en el aula 404.

4.2 Evaluación de configuraciones seleccionadas en el Aula 405

Esta sección presenta los resultados obtenidos al aplicar, sobre el conjunto de datos del aula 405, las configuraciones de hiperparámetros previamente seleccionadas en el aula 404 para la tarea de detección de ocupación. El objetivo de este análisis es evaluar la consistencia del desempeño de los modelos en un entorno distinto y analizar su capacidad de generalización.

4.2.1 Análisis de desempeño para detección de ocupación - Aula 405

La evaluación de los modelos en el aula 405 se realizó bajo las configuraciones óptimas identificadas previamente: SVM (RBF, $C = 10$), RF (200 árboles) y MLP (50, 25). Los resultados confirman una alta fiabilidad en la detección de ocupación, manteniendo la consistencia de las métricas de *Accuracy*, *Recall*, *Specificity*, AUC y *F1-score*.

El modelo SVM exhibió un desempeño estable con un *Accuracy* de 0.953 y un *F1-score* de 0.901. Según se observa en la matriz de confusión (Figura E.1), el algoritmo posee una adecuada capacidad discriminativa, limitando sus errores a confusiones puntuales que no comprometen la tendencia general de detección. Por su parte, el modelo RF se consolidó como la configuración de mejor desempeño global en este escenario; la Figura E.2 evidencia una alta tasa de clasificaciones

correctas y una reducción crítica de falsos positivos, lo que se traduce en valores superiores de *Specificity* y AUC.

En cuanto al modelo MLP, aunque presenta resultados competitivos (*Accuracy* de 0.954 y AUC de 0.982), su rendimiento promedio es marginalmente inferior al de RF debido a una mayor dispersión de errores reflejada en su matriz de confusión (Figura E.3). Los detalles exhaustivos de estas matrices se encuentran disponibles en el Anexo 5.

Finalmente, la Figura F.9 integra el análisis mediante un gráfico de tipo *radar*, donde se ratifica que el modelo RF posee el equilibrio métrico más sólido y la mayor área de cobertura. Las variaciones observadas respecto al aula 404 se atribuyen a la distribución intrínseca de los datos y los patrones de uso del aula 405, un hallazgo que subraya la importancia de la validación cruzada en entornos reales y refuerza la robustez del enfoque experimental adoptado para garantizar la generalización del algoritmo.

4.2.2 Análisis de desempeño para estimación de ocupación - Aula 405

En esta fase se validó la capacidad de generalización de las configuraciones óptimas identificadas previamente: SVM (RBF, $C = 10$), RF (200 árboles) y MLP (200, 100, 50). La evaluación se fundamenta en el análisis de métricas promediadas y en la comparativa visual del gráfico de radar (Figura F.9, Anexo 5), herramientas que permiten determinar la robustez de los modelos ante la variabilidad de los datos en un entorno distinto al de entrenamiento inicial.

El modelo Random Forest (RF) se consolidó como la arquitectura con mejor desempeño global, alcanzando un *Accuracy* de 0.722 y un AUC de 0.909. Según se observa en su matriz de confusión (Figura E.5), RF presenta la mayor concentración de aciertos en la diagonal principal, lo que demuestra una capacidad superior para discriminar entre niveles de ocupación multiclase.

Por su parte, el modelo MLP exhibió un desempeño intermedio con un *Accuracy* de 0.602 y un AUC de 0.834. Aunque logra capturar la tendencia general de la ocupación, la matriz de confusión (Figura E.6) revela una dispersión de errores concentrada en clases adyacentes, lo que afecta directamente su *Recall* y *F1-score*. Finalmente, el modelo SVM presentó las mayores limitaciones para la tarea multiclase, con un *Accuracy* de 0.536 y dificultades críticas para identificar

correctamente las clases de ocupación (Figura E.4).

En conclusión, los resultados en el aula 405 confirman la superioridad de RF en tareas de estimación compleja. La transición de los errores hacia clases contiguas en MLP y SVM, en lugar de errores aleatorios, valida la coherencia ordinal del modelo frente a la naturaleza gradual del fenómeno estudiado, aunque subraya la necesidad de arquitecturas de ensamble para maximizar la precisión en escenarios de alta variabilidad.

4.2.3 Análisis de desempeño final utilizando Radar Chart Aula 405

La Figura F.10 (Anexo 5) presenta el desempeño comparativo de los modelos mediante un gráfico de radar de métricas normalizadas. Este análisis evidencia que el modelo RF ofrece el comportamiento más equilibrado y la mayor cobertura de área, superando la estabilidad de SVM y MLP, cuyos desempeños son más limitados y dependientes de métricas específicas ante los patrones de ocupación del aula 405.

En conclusión, el modelo RF se consolida como la alternativa más robusta para la estimación de ocupación. Su superioridad técnica radica en la consistencia demostrada en ambos entornos (404 y 405), su capacidad de generalización y su eficacia para manejar la variabilidad de los datos en condiciones reales. Si bien SVM y MLP son competitivos en casos puntuales, su sensibilidad a las características del aula refuerza la selección de RF como el modelo óptimo para la discusión global en el siguiente capítulo.

4.3 Validación Comparativa: Random Forest vs. Estado del Arte (XGBoost)

Para ratificar la idoneidad de la arquitectura seleccionada, se contrastó el desempeño del modelo propuesto (*Random Forest*) frente a *Extreme Gradient Boosting* (XGBoost), algoritmo referente en el estado del arte para datos tabulares. El estudio se realizó bajo condiciones idénticas de preprocesamiento y partición de datos. Los resultados obtenidos se encuentran en el Anexo 5.

5 Conclusiones

Esta investigación validó el uso de variables ambientales y eléctricas no intrusivas como una alternativa eficaz y de bajo costo para la gestión de la ocupación en entornos educativos. A través de un análisis sistemático de aprendizaje supervisado con datos reales del edificio E3T de la UIS, se concluye que es viable inferir patrones de aforo con alta precisión.

En la tarea de detección de ocupación, se determinó que el modelo RF es la arquitectura más robusta. A diferencia de SVM y MLP, RF logró el equilibrio óptimo entre sensibilidad y especificidad, manteniendo un desempeño consistente en distintas condiciones ambientales. Esta estabilidad es crítica para sistemas de gestión energética, ya que minimiza falsos negativos que podrían comprometer el confort térmico o falsos positivos que derivarían en desperdicio energético.

Respecto a la estimación del nivel de ocupación, los resultados confirman un incremento en la complejidad predictiva; no obstante, el modelo RF superó consistentemente a los demás enfoques con un *Accuracy* de hasta 0.722 y un AUC de 0.909. Se observó que las imprecisiones se concentran en las fronteras de clases adyacentes, lo cual valida la coherencia ordinal del modelo. Las limitaciones en los niveles extremos de ocupación no se atribuyen a fallas algorítmicas, sino a la baja representatividad estadística de dichos eventos en el entorno académico real, un factor común en problemas de clasificación con datos desbalanceados.

La capacidad de generalización del sistema se ratificó mediante dos pruebas críticas: la evaluación y validación con los datos del aula 405 y la comparativa frente a técnicas de *gradient boosting* (XGBoost). El hecho de que el modelo RF mantuviera su eficacia en un entorno físico distinto sin reajuste de hiperparámetros, y que superara en estabilidad a XGBoost ante el ruido de los sensores, consolida al *bagging* como la estrategia de ensamble más apta para este dominio de aplicación.

Los resultados indican que la implementación de modelos de aprendizaje supervisado, particularmente RF, permite una estimación confiable de la ocupación, constituyendo un insumo técnico para estrategias de control dinámico en sistemas HVAC e iluminación orientadas a la optimización energética.

Referencias

ASHRAE. (2021). *ASHRAE handbook: HVAC applications*. American Society of Heating, Refrigerating y Air-Conditioning Engineers.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Candanedo, L. M., & Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy and Buildings*, *112*, 28-39. https://doi.org/10.1016/j.enbuild.2015.11.071

Chaudhari, P., Xiao, Y., Cheng, M. M., & Li, T. (2024). Fundamentals, algorithms, and technologies of occupancy detection for smart buildings using IoT sensors. *Sensors*, *24*(7), 2123. https://doi.org/10.3390/s24072123

Chen, J., Jiang, C., & Xie, L. (2018). Environmental sensing and neural networks for occupancy detection in smart buildings. *Applied Energy*, *225*, 1036-1049. https://doi.org/10.1016/j.apenergy.2018.05.039

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Chitnis, S., Somu, N., & Kowli, A. (2025). Occupancy estimation with environmental sensors: The possibilities and limitations. *Energy and Built Environment*, *6*(1), 96-108. https://doi.org/10.1016/j.enbenv.2023.09.003

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Decreto 2407, Colombia (2024).

Gunay, H. B., O'Brien, W., & Beausoleil-Morrison, I. (2019). Improving occupancy detection performance using semantic contextual features. *Building and Environment*, *156*, 215-228. https://doi.org/10.1016/j.buildenv.2019.04.033

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2.ª ed.). Springer.

Haykin, S. (2009). *Neural Networks and Learning Machines* (3.ª ed.). Prentice Hall.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3.ª ed.). Wiley.

International Energy Agency. (2023). *World energy outlook 2023*. IEA.

Ley 1715, Colombia (2014).

Li, X., & Fan, Y. (2020). Classroom occupancy detection using environmental sensors: A machine learning approach. *Energy and Buildings*, *224*, 110238. https://doi.org/10.1016/j.enbuild.2020.110238

Mena, A. R., Ceballos, H. G., & Alvarado-Uribe, J. (2022). Measuring indoor occupancy through environmental sensors: A systematic review on sensor deployment. *Sensors*, *22*(10), 3770. https://doi.org/10.3390/s22103770

Mena-Martinez, A., Alvarado-Uribe, J., Molino-Minero-Re, E., & Ceballos, H. G. (2024). Indoor occupancy monitoring using environmental feature fusion and semi-supervised machine learning models. *Journal of Building Performance Simulation*, *17*(6), 695-717. https://doi.org/10.1080/19401493.2024.2399053

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mohammadabadi, A., Rahnama, S., & Afshari, A. (2022). Indoor occupancy detection based on environmental data using CNN-XGBoost model: Experimental validation in a residential building. *Sustainability*, *14*(21), 14644. https://doi.org/10.3390/su142114644

Ortega-Diaz, L., Jaramillo-Ibarra, J., & Osma-Pinto, G. (2025). Estimation of the air conditioning energy consumption of a classroom using machine learning in a tropical climate. *Frontiers in Big Data*, *8*, 1520574. https://doi.org/10.3389/fdata.2025.1520574

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106. https://doi.org/10.1007/BF00116251

Rashid, K., Beg, M., & Al-Zahrani, A. (2018). Occupancy detection in buildings using machine learning techniques: A review. *Journal of Building Engineering*, *15*, 181-200. https://doi.org/10.1016/j.jobe.2017.11.020

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. United Nations. https://www.un.org/sustainabledevelopment/development-agenda/

United Nations Environment Programme. (2023). *Global status report for buildings and construction 2023*. UNEP. https://www.unep.org/resources/report/global-status-report-buildings-and-construction-2023

United Nations Framework Convention on Climate Change. (2015). *Paris Agreement*. United Nations. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

U.S. Department of Energy. (2022). *Energy use in schools: Building technologies program*. DOE.

Zeleny, M., Mach, P., & Mizera, T. (2024). Detection of room occupancy in smart buildings. *Radioengineering*, *33*(3), 432-441. https://doi.org/10.13164/re.2024.0432

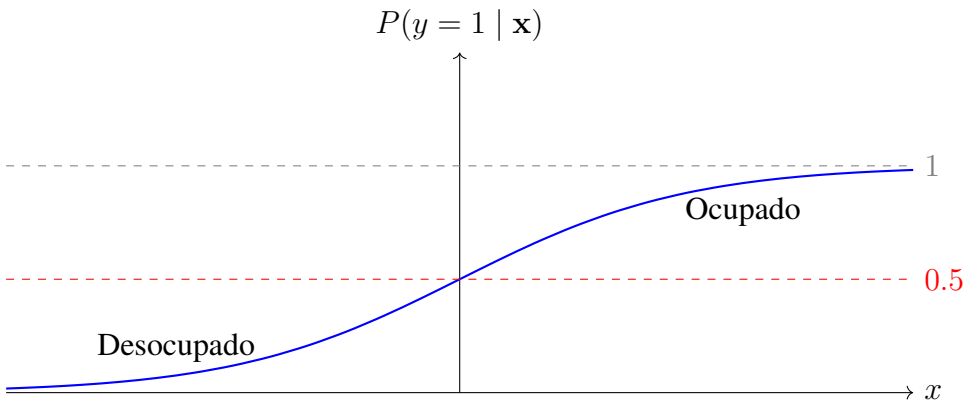
Zhang, T., & Jain, R. (2019). Environmental sensing for occupancy detection: A survey. *Sensors*, *19*(17), 3775. https://doi.org/10.3390/s19173775

Apéndice A. Técnicas supervisadas utilizadas para inferir ocupación

A.1 Regresión logística

Figura A.1

Función sigmoide y umbral de decisión en regresión logística.

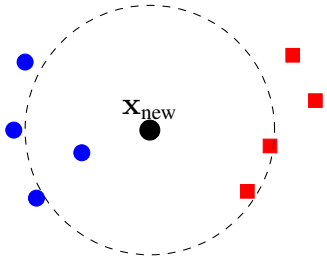


La Figura A.1 muestra una representación conceptual de la función sigmoide utilizada en la regresión logística, la cual transforma la combinación lineal de las variables de entrada en una probabilidad acotada entre 0 y 1, facilitando la interpretación del resultado del clasificador.

A.2 K-Nearest Neighbors

Figura A.2

Principio de clasificación del algoritmo KNN.



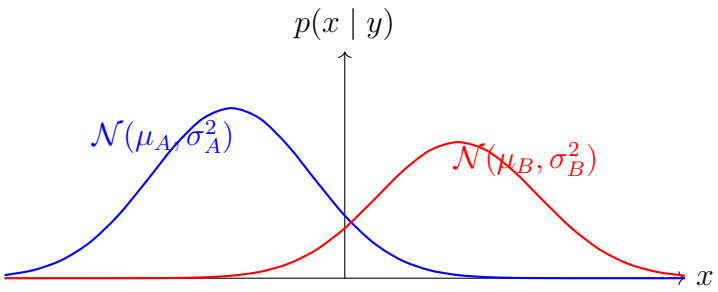
La Figura A.2 ilustra de manera conceptual el principio de funcionamiento del clasificador KNN, donde un nuevo punto es clasificado en función de la clase predominante entre sus vecinos

más cercanos en el espacio de características.

A.3 Naïve Bayes

Figura A.3

Funciones de densidad condicional en Naïve Bayes.

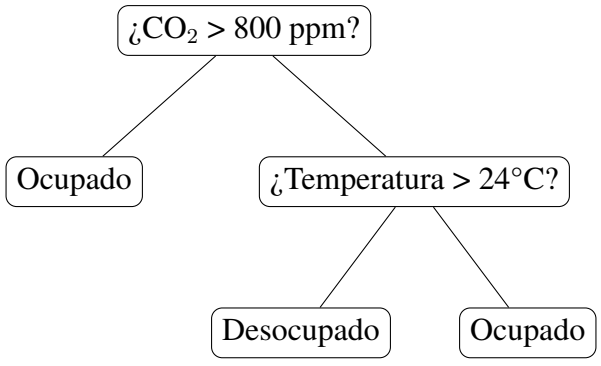


La Figura A.3 ilustra el principio de funcionamiento del clasificador Naïve Bayes, donde cada clase de ocupación es representada mediante una distribución probabilística independiente asociada a una variable característica.

A.4 Árboles de decisión

Figura A.4

Estructura conceptual de un árbol de decisión.



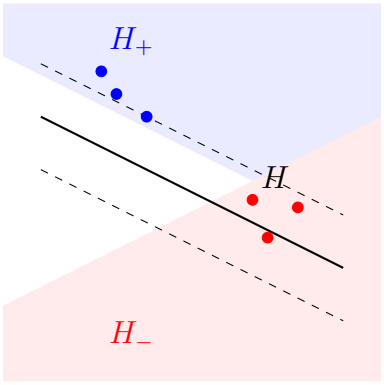
La Figura A.4 ilustra la estructura conceptual de un árbol de decisión aplicado a la detección de ocupación, donde variables ambientales como la concentración de CO_2 y la temperatura

interior se emplean para inferir el estado del aula mediante reglas jerárquicas de fácil interpretación.

A.5 Máquinas de Soporte Vectorial

Figura A.5

Clasificador SVM lineal y margen máximo.

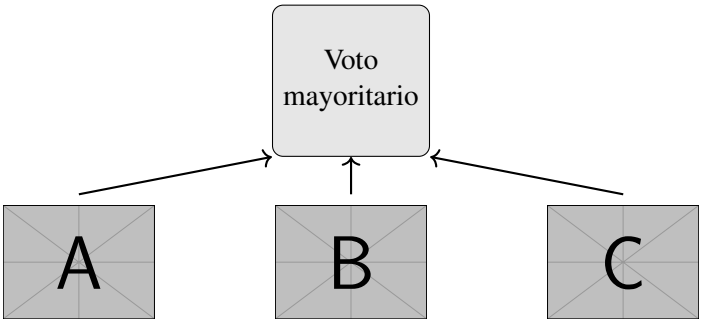


La Figura A.5 ilustra el principio de funcionamiento de un clasificador SVM lineal. El hiperplano H representa la frontera de decisión definida por el modelo, la cual separa las observaciones pertenecientes a diferentes clases. Las rectas H_+ y H_-, paralelas a H, corresponden a los márgenes positivo y negativo, respectivamente, y delimitan la región de separación máxima entre las clases. Las observaciones que se encuentran sobre estas rectas reciben el nombre de vectores de soporte y son las que determinan de manera directa la posición y orientación del hiperplano óptimo.

A.6 Random Forest

Figura A.6

Funcionamiento general del algoritmo Random Forest.

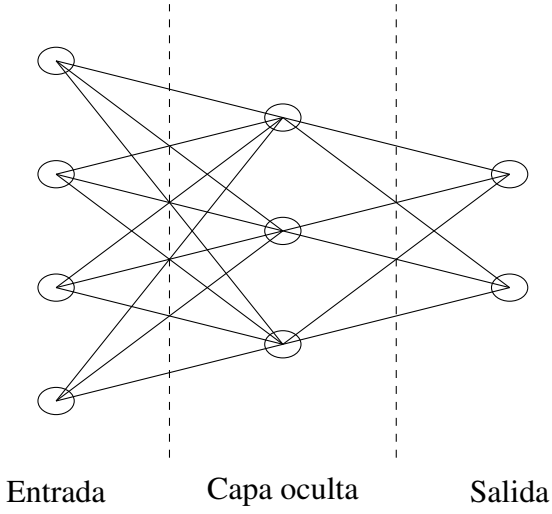


La Figura A.6 ilustra de manera conceptual el funcionamiento de Random Forest, en el que múltiples Random Forest generan predicciones independientes que posteriormente se combinan mediante un proceso de votación para inferir el estado de ocupación del espacio.

A.7 Multi-layer Perceptron

Figura A.7

Arquitectura general de un Multi-Layer Perceptron (MLP).



La Figura A.7 muestra una representación esquemática de la arquitectura básica de un MLP, ilustrando el flujo de información desde las variables de entrada, a través de una o más capas

ocultas, hasta la salida del modelo.

A.8 Comparación de las técnicas supervisadas

Tabla A.1

Comparación de técnicas supervisadas para detección y estimación de ocupación

Técnica	**Linealidad**	**Desbalance entre clases**	**Desempeño**	**Ventajas**	**Limitaciones**
Regresión logística	Lineal	Impacto moderado; requiere ajuste de umbral o ponderación	Moderado	Alta interpretabilidad; bajo costo computacional; fácil implementación	Limitada para modelar relaciones no lineales; desempeño reducido en escenarios complejos
k-Nearest Neighbors (k-NN)	No lineal	Alto impacto; sesgo hacia la clase mayoritaria	Moderado	Flexible ante distribuciones complejas; no requiere entrenamiento explícito	Alto costo computacional en inferencia; sensible al ruido y a la dimensionalidad
Naïve Bayes	Lineal	Alto impacto; probabilidad dominada por la clase mayoritaria	Bajo	Implementación sencilla; rápido; útil como referencia inicial	Suposición fuerte de independencia entre variables; desempeño limitado
Árbol de decisión	No lineal	Impacto moderado; depende del criterio de división	Moderado	Interpretabilidad; modela relaciones no lineales	Propenso al sobreajuste; inestabilidad ante variaciones en los datos
Máquinas de Soporte Vectorial (SVM)	No lineal	Bajo impacto; permite ponderación de clases	Alto	Buen desempeño en alta dimensión; márgenes robustos	Sensibilidad a hiperparámetros; mayor costo computacional
Random Forest (RF)	No lineal	Bajo impacto; combinación de múltiples árboles reduce sesgo	Alto	Reduce el sobreajuste; desempeño consistente; maneja relaciones complejas	Menor interpretabilidad; mayor costo computacional
Multi-Layer Perceptron (MLP)	No lineal	Impacto moderado; depende de la función de pérdida	Alto	Capacidad para aprender patrones complejos; arquitectura flexible	Requiere más datos; ajuste y entrenamiento complejos

Técnica	**Linealidad**	**Desbalance entre clases**	**Desempeño**	**Ventajas**	**Limitaciones**
XGBoost	No lineal	Impacto moderado; puede manejar desbalance mediante ponderación de clases	Alto	Alto desempeño predictivo; regularización integrada; robusto a ruido	Sensible a hiperparámetros; mayor complejidad computacional

Apéndice B. Caso de estudio - Aulas E3T

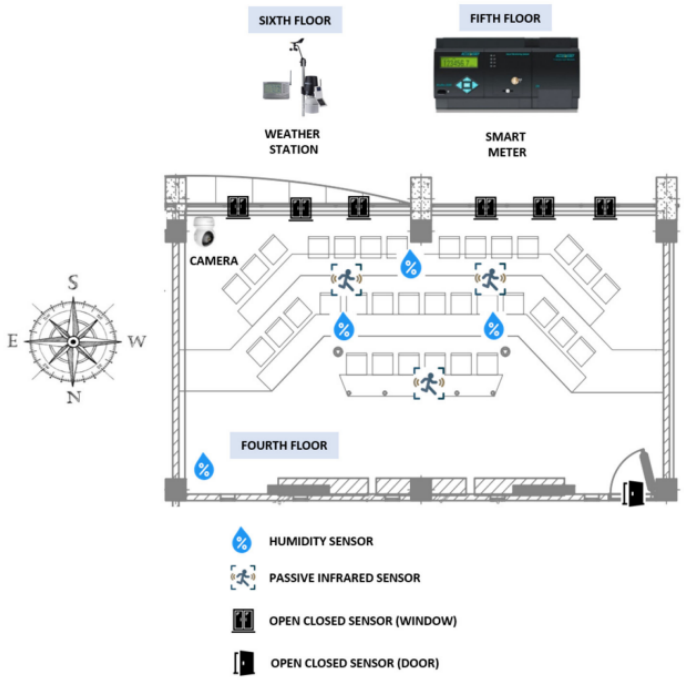
B.1 Imagenes Aula y ubicación de los sensores

Figura B.1

Vista general del caso de estudio: Aula. Fuente: tomada de Ortega-Diaz et al. (2025).

Figura B.2

Vista general del caso de estudio: aula y ubicación de los sensores. Fuente: tomada de Ortega-Diaz et al. (2025).



B.2 Variables medidas

Tabla B.1

Variables medidas en el conjunto de datos utilizado para el análisis.

Categoría	Variable	Descripción	Tipo
Energéticas	Active Power (W)	Potencia activa consumida por equipos.	Continua
	Energy Consumption (Wh)	Energía acumulada consumida en el intervalo.	Continua
Amb. exteriores	T. Outdoor (°C)	Temperatura exterior.	Continua
	H. Outdoor (%)	Humedad exterior.	Continua
	Dew Point (°C)	Punto de rocío.	Continua
	Wind Speed (m/s)	Velocidad del viento.	Continua
	Wind Direction (°)	Dirección del viento.	Continua
	Heat Index (°C)	Índice de calor.	Continua
	Atmospheric Pressure (hPa)	Presión atmosférica.	Continua
	Rain Rate (mm/h)	Intensidad de lluvia.	Continua
	Solar Radiation (W/m^2)	Radiación solar.	Continua
	UV Index	Radiación UV.	Continua
	Cool Degree Days	Índice de refrigeración.	Continua
Internas del aula	Door Status	Puerta abierta/cerrada.	Binaria (0,1)
	Windows Status	Ventanas abiertas/cerradas.	Binaria (0,1)
	T. Indoor (°C)	Temperatura interior.	Continua
	H. Indoor (%)	Humedad interior.	Continua
	Motion	Movimiento detectado.	Binaria (0,1)
Ocupación	Occupant Number	Conteo real de personas.	Discreta $(0, \ldots, N)$
	Computer Number	Computadores en uso.	Discreta $(0, \ldots, M)$
	Occupancy	Estado de ocupación.	Binaria (0,1)

Tabla B.2

Variables finales utilizadas para el entrenamiento de los modelos

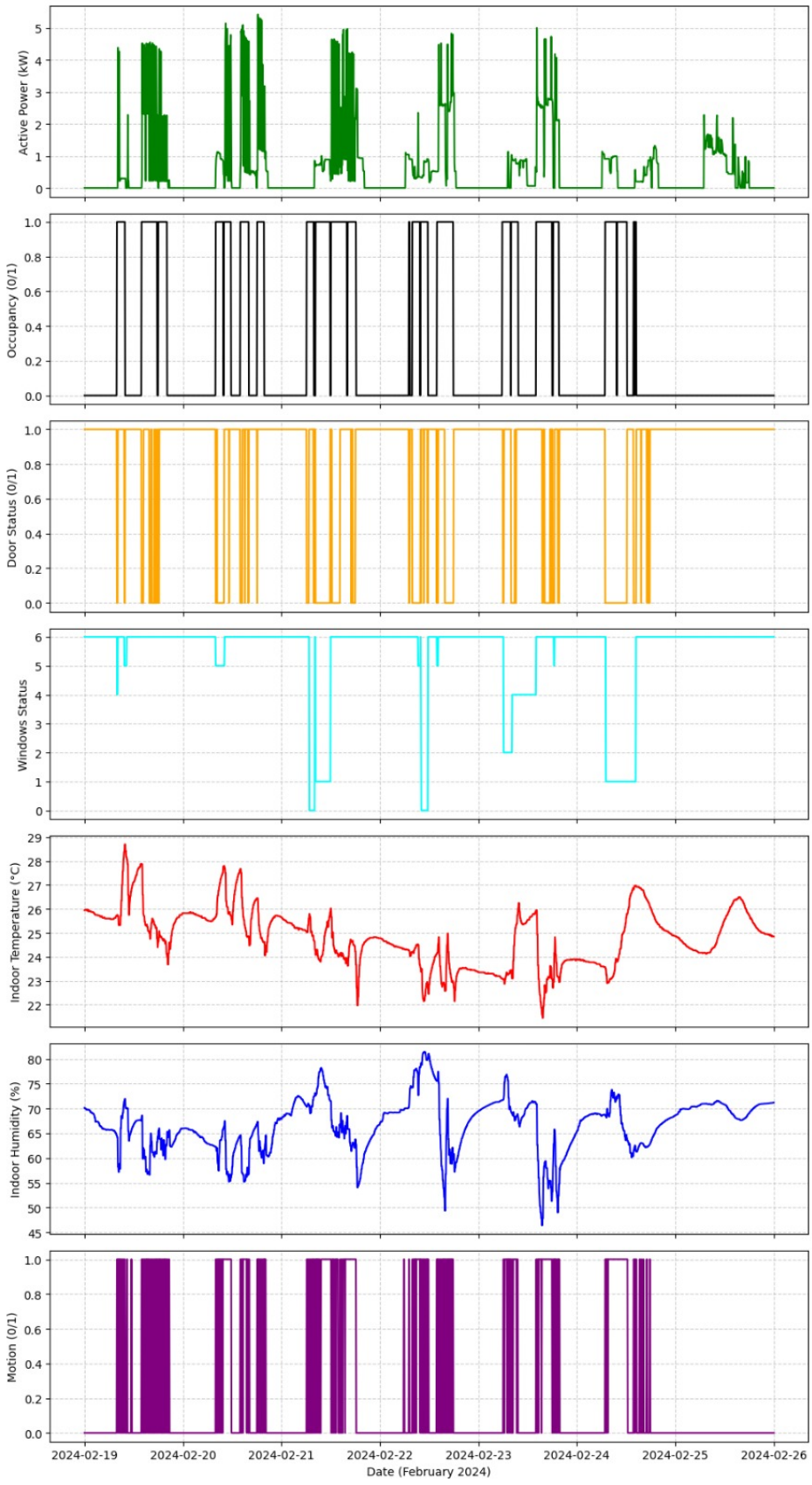
Símbolo	Variable	Tipo
T_{int}	Temperatura interior	Continua
HR	Humedad relativa	Continua
CO_2	Concentración de CO_2	Continua
L	Iluminación	Continua
D	Estado de puertas	Binaria
V	Estado de ventanas	Binaria

B.3 Comportamiento semanal de las variables empleadas

La Figura B.3 ilustra la evolución temporal de las variables de entrenamiento durante una semana representativa (19–26 de febrero de 2024). Estos registros permiten analizar los patrones diarios de las condiciones ambientales, el uso del aula y la presencia de ocupantes, así como las interdependencias temporales entre estas magnitudes.

Figura B.3

Comportamiento semanal de las variables empleadas.



En este anexo se detallan los procedimientos secuenciales aplicados para la preparación de los datos y la configuración experimental.

Tabla B.3

Resumen del proceso de preprocesamiento aplicado al conjunto de datos

Paso	**Descripción**
1	Cargar la base de datos original D.
2	Eliminar registros inválidos, inconsistentes o duplicados.
3	Codificar variables categóricas (estado de puertas y ventanas).
4	Normalizar variables numéricas mediante *StandardScaler*.
5	Seleccionar las variables relevantes para el análisis.
6	Generar el conjunto de datos preprocesado D'.

Apéndice C. Distribución de clases y desbalance del conjunto de datos

Este anexo presenta la distribución porcentual de las clases utilizadas en las tareas de detección de ocupación y estimación de rangos de ocupación para las aulas 404 y 405 de la E3T–UIS. Esta información permite cuantificar el grado de desbalance presente en los conjuntos de datos empleados durante el entrenamiento y evaluación de los modelos.

Distribución de clases para detección de ocupación

La Tabla C.1 resume la distribución de las clases binarias (ocupado y desocupado) para ambas aulas. Se observa un desbalance significativo a favor de la clase sin ocupación, lo cual es consistente con el patrón real de uso de las aulas académicas.

Tabla C.1

Distribución porcentual de clases para la detección de la ocupación

Aula	Clase	Conteo	Porcentaje (%)
404	0 (sin ocupación)	17098	79.96
404	1 (con ocupación)	4284	20.04
405	0 (sin ocupación)	14158	81.13
405	1 (con ocupación)	3294	18.87

Distribución de clases para estimación de ocupación

La Tabla C.2 presenta la distribución porcentual de los rangos de ocupación definidos para la tarea multiclase. Se evidencia un predominio del rango R0 (sin ocupación), así como una reducción progresiva del número de muestras en rangos superiores de ocupación.

Tabla C.2

Distribución porcentual de clases para la estimación de la ocupación

Rango	Personas	Aula 404 (%)	Aula 405 (%)
R0	0	79.96	81.13
R1	1–7	7.39	6.44
R2	8–14	4.78	4.03
R3	15–21	4.29	4.11
R4	22–28	3.01	2.59
R5	29–35	0.57	1.70

Apéndice D. Tablas de resultados experimentales

Tabla D.1

Resultados de SVM para la detección de ocupación en el aula 404

Configuración	Acc.	Rec.	Spec.	AUC	F1	Tiempo (s)	Prom.
Lineal $C = 1$	0.942	0.953	0.940	0.982	0.869	12.20	0.937
Lineal $C = 10$	0.942	0.953	0.940	0.982	0.869	32.93	0.937
Polinomial $C = 1$	0.960	0.939	0.966	0.974	0.905	25.48	0.949
Polinomial $C = 10$	0.960	0.939	0.965	0.973	0.904	109.85	0.948
RBF $C = 1$	0.958	0.964	0.957	0.975	0.902	10.41	0.951
RBF $C = 10$	**0.960**	**0.967**	**0.958**	**0.977**	**0.906**	**55.79**	**0.954**

Tabla D.2

Resultados de RF para la detección de ocupación en el aula 404

Árboles	Acc.	Rec.	Spec.	AUC	F1	Tiempo (s)	Prom.
50	0.974	0.955	0.979	0.993	0.937	2.23	0.968
100	0.975	0.956	0.980	0.993	0.939	2.66	0.969
200	**0.976**	**0.960**	**0.980**	**0.993**	**0.941**	**4.62**	**0.970**
300	0.975	0.958	0.979	0.994	0.938	6.07	0.969
400	0.975	0.959	0.980	0.994	0.940	8.54	0.970

Tabla D.3

Resultados de MLP para la detección de ocupación en el aula 404

Arquitectura	Acc.	Rec.	Spec.	AUC	F1	Tiempo (s)	Prom.
(50,25)	**0.969**	**0.932**	**0.978**	**0.990**	**0.922**	**4.18**	**0.958**
(100,50)	0.968	0.935	0.976	0.987	0.921	4.25	0.957
(100,50,25)	0.964	0.928	0.973	0.985	0.912	9.88	0.952
(200,100)	0.965	0.928	0.974	0.986	0.914	8.38	0.953
(200,100,50)	0.967	0.920	0.979	0.987	0.918	9.88	0.954

Tabla D.4

Resultados de SVM para la estimación de ocupación en el aula 404

Configuración	Acc.	Rec.	AUC	F1	Tiempo (s)	Prom.
Lineal $C = 1$	0.356	0.275	0.643	0.262	3.79	0.472
Lineal $C = 10$	0.357	0.271	0.640	0.263	13.76	0.471
Polinomial $C = 1$	0.407	0.309	0.700	0.285	7.44	0.508
Polinomial $C = 10$	0.431	0.333	0.707	0.319	9.32	0.527
RBF $C = 1$	0.442	0.342	0.732	0.329	8.56	0.538
RBF $C = 10$	**0.490**	**0.389**	**0.758**	**0.382**	**5.46**	**0.576**

Tabla D.5

Resultados de RF para la estimación de ocupación en el aula 404

Árboles	Acc.	Rec.	AUC	F1	Tiempo (s)	Prom.
50	0.658	0.583	0.876	0.597	0.61	0.724
100	0.656	0.586	0.881	0.597	1.94	0.725
200	**0.660**	**0.588**	**0.882**	**0.604**	**1.67**	**0.728**
300	0.656	0.586	0.883	0.600	5.39	0.726
400	0.657	0.587	0.883	0.600	4.68	0.727

Tabla D.6

Resultados de MLP para la estimación de ocupación en el aula 404

Arquitectura	Acc.	Rec.	AUC	F1	Tiempo (s)	Prom.
(50,25)	0.440	0.348	0.725	0.343	6.96	0.541
(100,50)	0.401	0.311	0.695	0.301	8.08	0.509
(100,50,25)	0.421	0.325	0.715	0.314	8.08	0.524
(200,100)	0.466	0.363	0.751	0.351	8.08	0.557
(200,100,50)	**0.490**	**0.391**	**0.768**	**0.384**	**8.08**	**0.579**

Tabla D.7

Resultados de los modelos seleccionados para la detección de ocupación en el aula 405

Modelo	**Configuración**	Acc.	Rec.	Spec.	AUC	F1	Promedio
SVM	RBF $C = 10$	0.953	0.952	0.954	0.967	0.901	0.945
RF	**200 árboles**	**0.957**	**0.919**	**0.968**	**0.991**	**0.905**	**0.948**
MLP	(50, 25)	0.954	0.908	0.967	0.982	0.897	0.941

Tabla D.8

Resultados de los modelos seleccionados para la estimación de ocupación en el aula 405

Modelo	**Configuración**	Acc.	Rec.	AUC	F1	Promedio
SVM	RBF $C = 10$	0.536	0.428	0.787	0.446	0.6136
RF	**200 árboles**	**0.722**	**0.685**	**0.909**	**0.698**	**0.7878**
MLP	(200, 100, 50)	0.602	0.526	0.834	0.536	0.6778

Apéndice E. Matrices de confusión

Figura E.1

Matriz de confusión del modelo SVM para la detección de ocupación (aula 405).

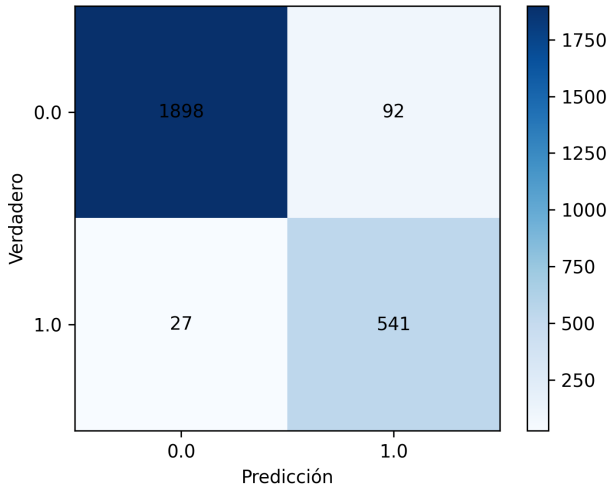


Figura E.2

Matriz de confusión del modelo RF para la detección de ocupación (aula 405).

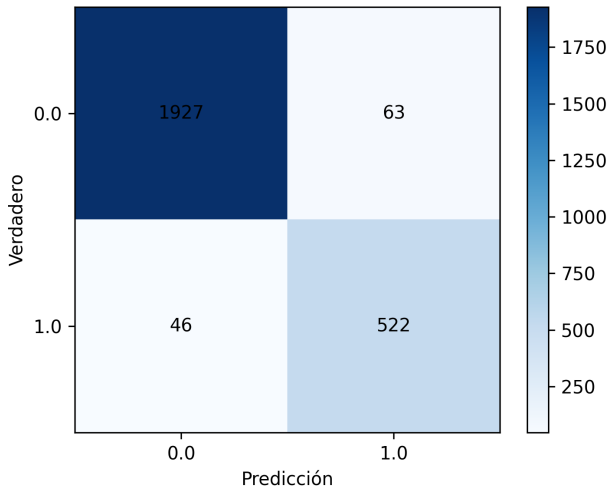


Figura E.3

Matriz de confusión del modelo MLP para la detección de ocupación (aula 405).

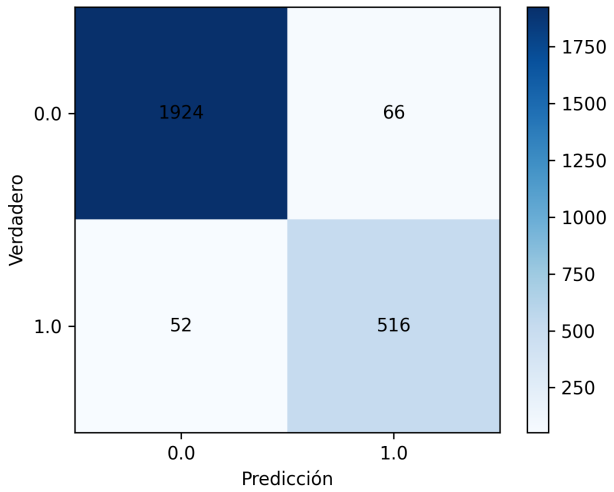


Figura E.4

Matriz de confusión del modelo SVM para la estimación de ocupación (aula 405).

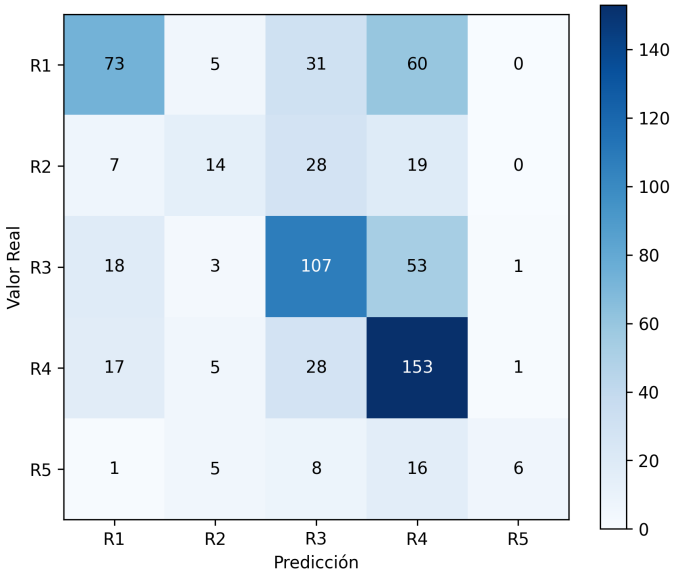


Figura E.5

Matriz de confusión del modelo RF para la estimación de ocupación (aula 405).

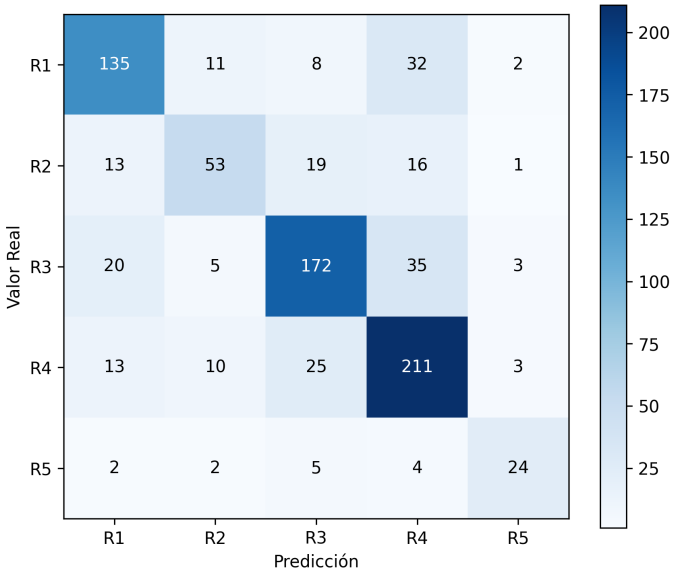
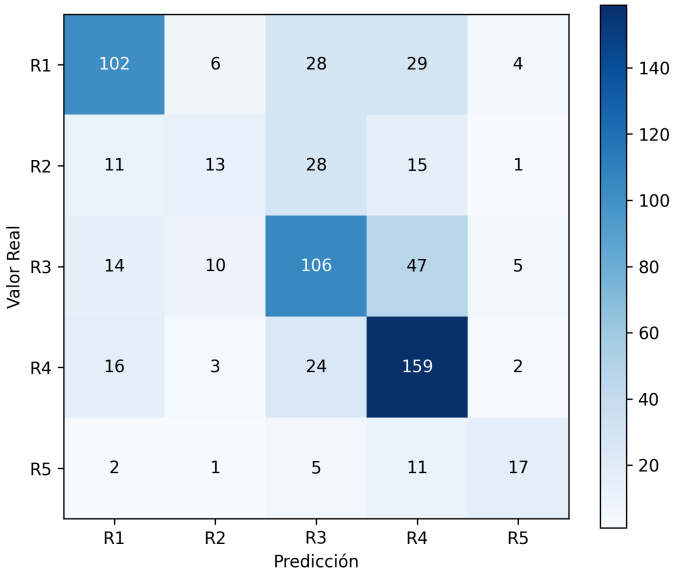


Figura E.6

Matriz de confusión del modelo MLP para la estimación de ocupación (aula 405).



Apéndice F. Desempeño detallado de modelos

Figura F.1

Desempeño promedio de las configuraciones SVM (Detección, aula 404).

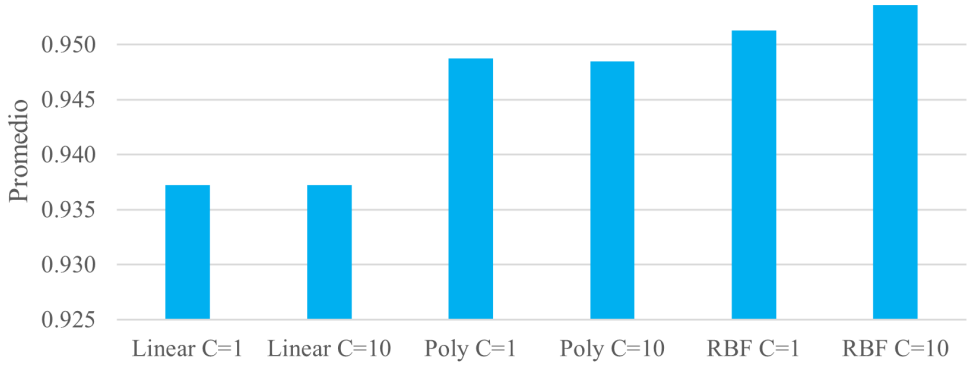


Figura F.2

Desempeño promedio de las configuraciones RF (Detección, aula 404).

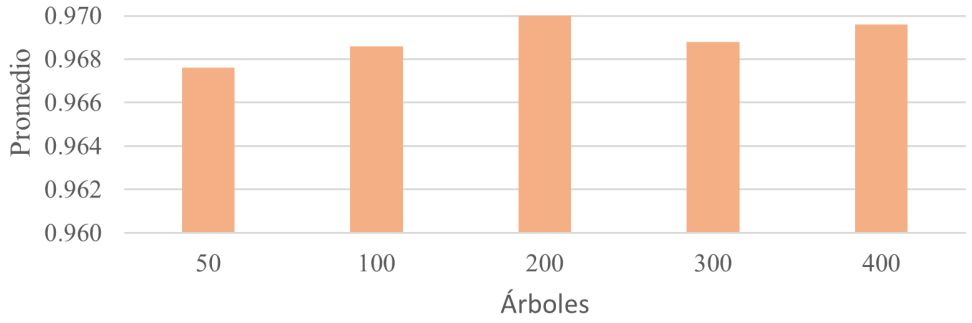


Figura F.3

Desempeño promedio de las configuraciones MLP (Detección, aula 404).

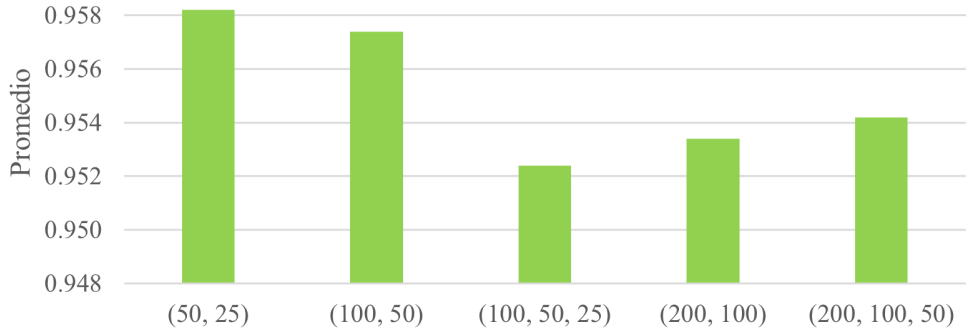


Figura F.4

Desempeño promedio de las configuraciones SVM (Estimación, aula 404).

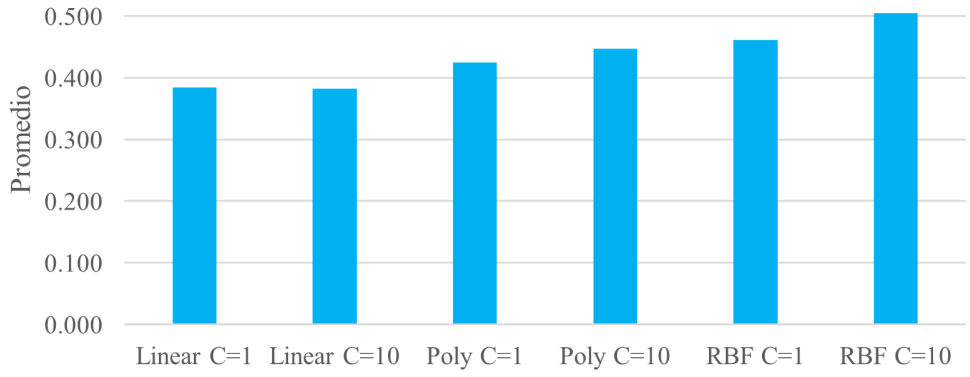


Figura F.5

Desempeño promedio de las configuraciones RF (Estimación, aula 404).

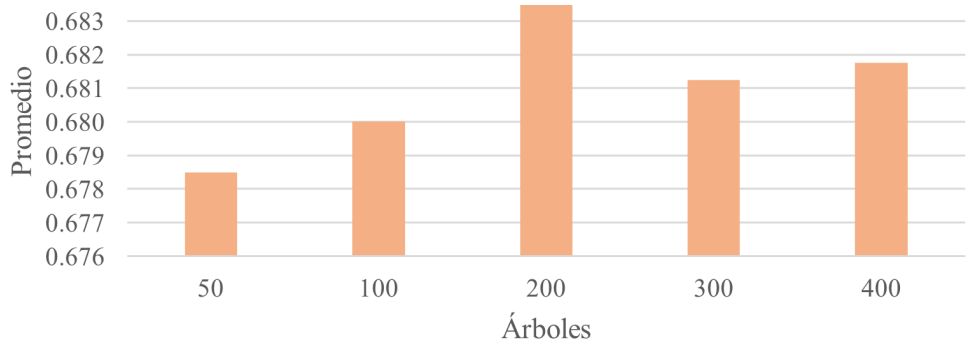
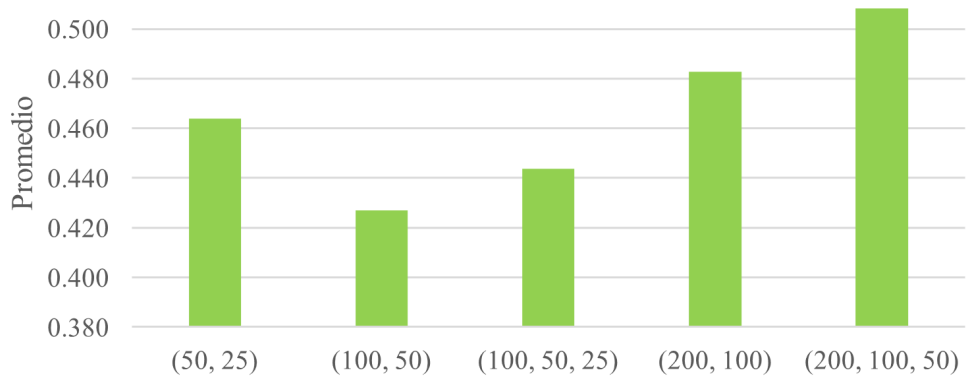


Figura F.6

Desempeño promedio de las configuraciones MLP (Estimación, aula 404).



Comparación global de algoritmos

Figura F.7

Comparación del desempeño de los algoritmos (Detección aula 404).

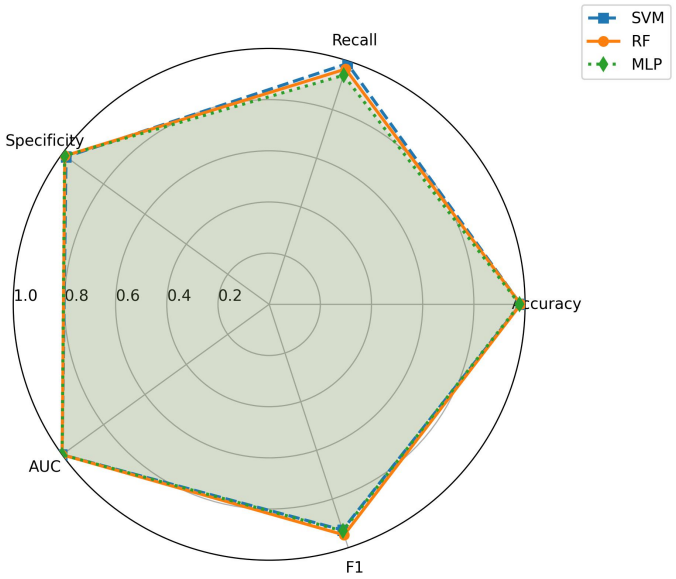


Figura F.8

Comparación del desempeño de los algoritmos (Estimación aula 404).

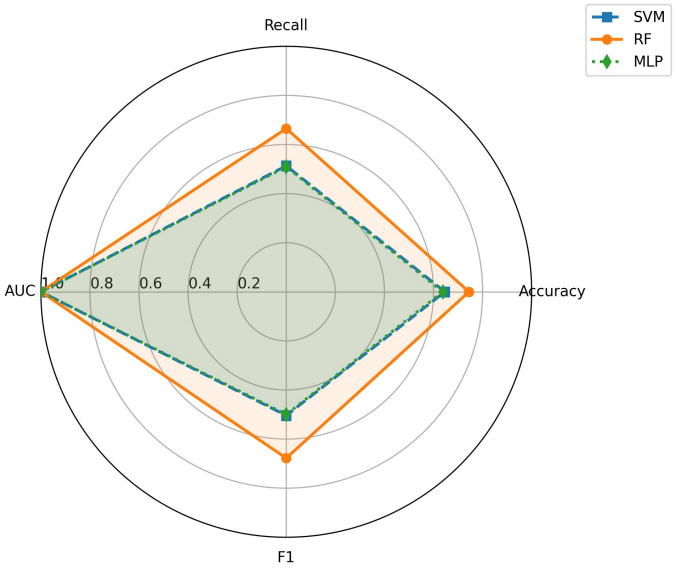


Figura F.9

Comparación del desempeño de los algoritmos (Detección aula 405).

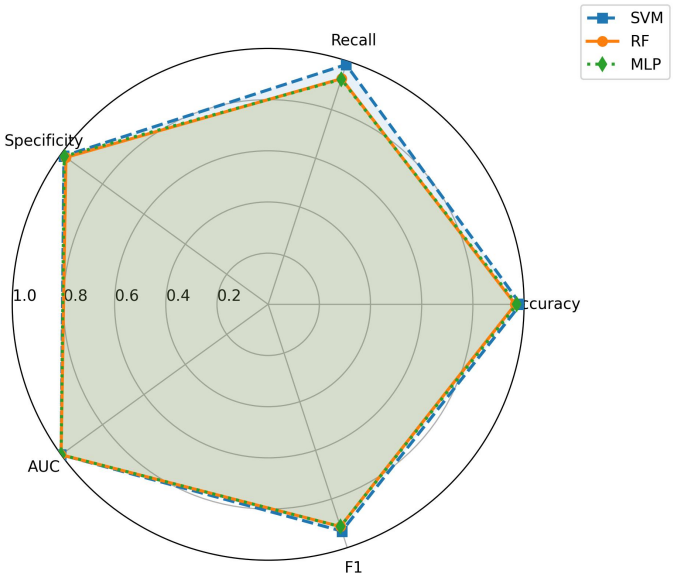
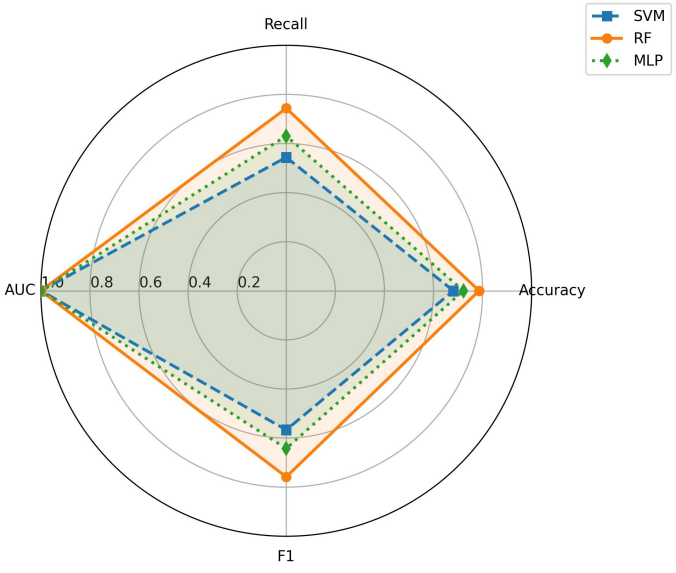


Figura F.10

Comparación del desempeño de los algoritmos en (Estimación aula 405).



Apéndice G. Validación con XGBoost

Detección de la ocupación

En la tarea de detección de ocupación, el modelo seleccionado (*RF* con 200 árboles) demostró una superioridad técnica consistente sobre XGBoost. La Tabla G.1 presenta el cruce de métricas exactas.

Tabla G.1

Comparativa de desempeño: Detección Binaria (Aula 404)

Modelo	Accuracy	Recall	Specificity	F1-Score	AUC
Random Forest (200 árboles)	**0.976**	**0.960**	**0.980**	**0.941**	**0.993**
XGBoost	0.973	0.946	0.980	0.934	0.993

Análisis de resultados:

- **Mejor Garantía de Confort:** El hallazgo más relevante es la diferencia en la sensibilidad (*Recall*). RF alcanzó un 96.0 %, superando al 94.6 % de XGBoost. Esto implica que el sistema propuesto tiene menor probabilidad de fallar detectando personas (falsos negativos), lo cual es crítico para evitar que el sistema de climatización se apague indebidamente cuando hay usuarios en el aula.

- **Eficiencia Equivalente:** Ambos modelos lograron una especificidad idéntica (98.0 %), garantizando el mismo nivel de ahorro energético al identificar correctamente los periodos de vacancia. Sin embargo, el mayor F1-Score de Random Forest (0.941 vs 0.934) confirma que es el modelo más equilibrado globalmente.

Estimación de la ocupación

Para la tarea de estimación de rangos de ocupación (multiclase), se comparó la capacidad de ambos algoritmos para discriminar entre los 5 niveles de ocupación positiva. Los resultados se detallan en la Tabla G.2.

Tabla G.2

Comparativa de desempeño: Estimación Multiclase (Aula 404)

Modelo	Accuracy	Recall (Macro)	F1-Score (Macro)	AUC (Macro)
Random Forest (200 árboles)	**0.660**	**0.588**	**0.604**	**0.882**
XGBoost	0.645	0.588	0.603	0.879

Análisis de la robustez: En esta tarea de mayor complejidad, RF mantiene su liderazgo frente a XGBoost:

1. **Precisión Global:** El modelo propuesto supera a XGBoost en exactitud ($66.0\,\%$ vs $64.5\,\%$). Esta diferencia valida que la estrategia de *Bagging* (promedio de árboles independientes) es más efectiva para generalizar en fronteras de decisión difusas —típicas de la estimación de aforo con variables ambientales— que la estrategia de *Boosting*, la cual tiende a ser más sensible al ruido de los sensores en clases solapadas.

2. **Estabilidad:** Aunque el *Recall* promedio es idéntico, la ligera ventaja en AUC (0.882 vs 0.879) sugiere que Random Forest ofrece una mejor separación probabilística entre las clases, lo que se traduce en predicciones más confiables para el sistema de control.

Conclusión del Estudio: La validación experimental confirma que RF es la arquitectura óptima para este sistema, superando al estado del arte (XGBoost) tanto en la precisión de la estimación como en la confiabilidad de la detección.