Secretaría de Planeación Distrital

Centro de Estudios Territoriales
Alcaldía de Cartagena

Identificación de procesos de Gentrificación Urbana: Modelo Random Forest Aplicado a Cartagena, Colombia

WORKING PAPERS CET

NÚMERO 01

Autores:

Guatecique Lidueña Laima

Gonzalez Agudelo Jose Daniel

IDENTIFICACIÓN DE PROCESOS DE GENTRIFICACIÓN URBANA: MODELO RANDOM FOREST APLICADO A CARTAGENA, COLOMBIA

Guatecique Lidueña Laima *

Gonzalez Agudelo Jose Daniel ♣

Resumen

Los avances recientes en herramientas para identificar procesos de gentrificación dentro de dinámicas urbanas han impulsado el uso de algoritmos de aprendizaje automático, capaces de ofrecer predicciones y análisis más precisos en los estudios sobre urbanización y transformación socioespacial. En este contexto, desarrollamos un modelo de Machine Learning basado en Random Forest para estimar la probabilidad de gentrificación en los barrios de Cartagena de Indias. El cambio urbano se evalúa mediante variables socioeconómicas extraídas de datos censales, complementadas con un indicador que clasifica la presencia de transformaciones en la dinámica urbana. El modelo predice el cambio entre 2005 y 2018 con una precisión (accuracy) del 91,4%. Los resultados evidencian una expansión de la gentrificación en la zona norte de la ciudad y en el centro, particularmente en el centro histórico. Se prevé que estos hallazgos ofrezcan a los responsables de políticas públicas una herramienta eficaz para identificar con mayor exactitud las áreas con mayor probabilidad de experimentar gentrificación. Esta capacidad predictiva puede orientar la formulación de intervenciones y estrategias que promuevan un desarrollo urbano más equitativo, especialmente en beneficio de las comunidades vulnerables afectadas por los procesos de transformación barrial.

Palabras clave: Cartagena, random forest, machine learning, gentrificación, dinámica urbana, variables socioeconómicas

* Asesora externa, Centro de Estudios Territoriales (CET), secretaria de Planeación Distrital, Alcaldía de Cartagena.

♣ Asesor externo, Centro de Estudios Territoriales (CET), secretaria de Planeación Distrital, Alcaldía de Cartagena.

WORKING PAPERS CET

Secretaría de Planeación Distrital

Centro de Estudios Territoriales
Alcaldía de Cartagena

IDENTIFICATION OF URBAN GENTRIFICATION PROCESSES: RANDOM FOREST MODEL APPLIED TO CARTAGENA, COLOMBIA

Guatecique Lidueña Laima *

Gonzalez Agudelo Jose Daniel ♣

Abstract

Recent advances in tools for identifying gentrification processes within urban dynamics have encouraged the use of machine learning algorithms, which offer more precise predictions and analyses in the study of urbanization and socio-spatial transformation. In this context, we developed a Machine Learning model based on Random Forest to estimate the probability of gentrification in the neighborhoods of Cartagena de Indias. Urban change is assessed using socioeconomic variables derived from census data, complemented by an indicator that classifies the presence of transformations in urban dynamics.The model predicts change between 2005 and 2018 with an accuracy of 91.4%. The results show an expansion of gentrification in the northern part of the city and in the central area, particularly in the historic center. These findings are expected to provide policymakers with an effective tool to more accurately identify areas with a higher likelihood of experiencing gentrification. This predictive capacity can guide the development of interventions and strategies that promote more equitable urban development, especially for vulnerable communities affected by neighborhood transformation processes.

Keywords: Cartagena, random forest, machine learning, gentrification, urban dynamics, socioeconomic variables

* External advisor, Center for Territorial Studies (CET), District Planning Secretary, Mayor's Office of Cartagena.

♣ External advisor, Center for Territorial Studies (CET), District Planning Secretary, Mayor's Office of Cartagena.

1.INTRODUCCIÓN

La gentrificación se refiere a la llegada de hogares con mayor movilidad social a vecindarios con un menor nivel socioeconómico (Salinas Arreortua, 2013). Esto presenta un cambio social y transformación urbana dentro de barrios de clases populares, que se ven enfrentados a desplazarse gracias a los nuevos residentes de clase media y alta, que traen consigo directa o indirectamente (es decir por cuenta propia o inversión privada de agentes inmobiliarios) mejoras en infraestructura y viviendas, aunado a un aumento de valor en el suelo (Villanueva & Vallbona, 2021). Este fenómeno de dinámica urbana se observa por un gran número de ciudades mundiales, principalmente aquellas que cuentan con un fuerte atractivo turístico.

A lo largo de los años se ha presentado un debate sobre el concepto de gentrificación, dado que estos procesos no siguen un reglas o pautas establecidas debido a que está sujeto a las dinámicas socio-espaciales y la resistencia de las poblaciones al cambio, sin embargo, presenta sus orígenes a mediados del siglo XX con el fin de explicar acontecimientos repetitivos en importantes ciudades anglosajonas que dejaban de ser centros con actividades productivas y pasaban a tener su economía basada en actividades financieras, convirtiéndose en espacios de inversión y especulación inmobiliaria (Castro et al., 2020; Perren & Cabezas, 2018). Descrito por primera vez en 1964 por la socióloga británica Ruth Glass en un análisis en la ciudad de Londres, donde residencias modestas de arrendamiento de corto plazo se convertían en alojamientos caros y lujosos, los cuales se iban expandiendo rápidamente hasta desembocar en desalojo masivo de los inquilinos obreros que permanecían allí originalmente (Salinas Arreortua, 2013).

En el contexto moderno se han desarrollado diversos estudios que buscan evidenciar los nuevos procesos de gentrificación vinculados al capitalismo de plataformas. Un trabajo reciente titulado “Ciudades y globalización: capitalismo de plataformas y gentrificación en Nueva York, Londres y Ciudad de México (2008–2023)” (Mejía, 2024), analiza la relación entre ambos fenómenos a partir del auge de los hospedajes de renta corta, especialmente a través de plataformas como Airbnb. Mediante un enfoque cualitativo, el estudio examina leyes locales, informes, acuerdos, propuestas de ley, ordenanzas, comunicaciones oficiales, censos y bases de datos con información sobre Airbnb. Los hallazgos plantean que la gentrificación contemporánea adquiere un carácter comercial, derivado de las transformaciones demográficas, urbanas y sociales con fines mercantiles, lo que ha generado la expulsión de habitantes de bajos ingresos ante el incremento en el costo de vida. El estudio demuestra que dichas ciudades utilizan a Airbnb como un mecanismo de gentrificación debido a sus características: alta escalabilidad de operaciones a nivel global, corporativización de la gobernanza de datos, implementación de un modelo de negocio híbrido que combina elementos de mercado y empresa, externalización de costos y riesgos de producción, uso de narrativas asociadas a la economía compartida y la captura de valor. En el caso de Nueva York, la empresa combinó estrategias de lobby, litigios, alianzas políticas y campañas mediáticas para legitimar una actividad inicialmente ilegal, promoviendo una imagen de intermediario neutral y socialmente responsable. En Londres, su influencia fue menor debido a que los políticos locales emplearon la “regulación desregulada” como herramienta de formalización, lo que permitió la expansión de la plataforma con escasa oposición social. En contraste, en la Ciudad de México, Airbnb

ejerció un fuerte poder estructural al establecer alianzas con el gobierno y con organismos internacionales como la UNESCO y la Organización Mundial del Turismo (OMT), impulsando políticas favorables bajo el discurso del turismo sostenible y el trabajo remoto, las cuales incluso influyeron en cambios en las políticas migratorias locales.

Otro estudio importante situado en el contexto latinoamericano es ¿Renovación sin gentrificación? Hacia un abordaje crítico de procesos urbanos excluyentes en América Latina. Casos en Buenos Aires (Lerena-Rongvaux, 2023), donde explica cómo la ciudad de Buenos Aires ha experimentado políticas de renovación urbana encaminadas a la valorización de su zona sur la Comuna 4, área rezagada por el Estado. La metodología utilizada fue cualicuantitativa bajo cuatro enfoques de dinámica: estructura sociohabitacional (usando indicadores demográficos, económicos y habitacionales), valorización del mercado inmobiliario (usando precio promedio absoluto por metro cuadrado por tipología inmobiliaria) y de suelo, políticas de renovación (considerando leyes distritales y planes urbanos) y finalmente, el tejido organizativo y comunitario (indicadores relacionales y de acción colectiva). La autora concluye el estudio sobre la gentrificación para medir la renovación urbana y desplazamiento poblacional tiene limitaciones en América latina, donde pueden existir procesos de reinversión sin expulsión inmediata de habitantes. Por ello, se propone la noción de “renovación urbana excluyente”, que permite anticipar desigualdades incluso sin desplazamientos evidentes. El estudio examina los efectos de los Distritos Económicos del Sur de Buenos Aires (Tecnológico y de las Artes), mostrando que ambos experimentan valorización del suelo, aunque con resultados distintos: en el Tecnológico hay valorización sin gran conflicto social, mientras que en el de las Artes la precariedad habitacional genera tensiones y desplazamientos. En conjunto, se evidencia un proceso de renovación urbana excluyente, impulsado por políticas públicas que favorecen la inversión sin mecanismos que protejan a los sectores más vulnerables.

El caso de Cartagena, focal de estudio de este artículo, representa un proceso de gentrificación no solo a nivel socioeconómico, sino también social e histórico, el efecto expulsión de la población de ingresos bajos de los barrios populares es notable, en el caso de Getsemaní, en alrededor de quince años, la población residente se redujo en un 80 %, pasando de casi 10.500 personas en 2005 a poco más de 2.300 en 2018, según El Diario (España, citado en Barrios Uribe, 2024) y en el 2025 contando tan solo con 448 habitantes (Universidad de Cartagena, 2025). Este drástico descenso refleja un profundo impacto social: los pocos habitantes que permanecen resisten al alza del valor del suelo y al avance de un turismo desregulado, mientras defienden colectivamente su derecho a seguir habitando el territorio. Para el ámbito latinoamericano y en especial ciudades como Cartagena la gentrificación ha sido estudiada como dinámica urbana impulsada tanto por capital inmobiliario como por políticas públicas que dejan en un plano secundario, e incluso marginales el impacto social en consecuencia de priorizar el desarrollo económico, utilizando como principal vía la turistificación, mejorando la calidad paisajística.

Para Cartagena también se aplica el fenómeno de gentrificación trasnacional, el cual incluye como premisas de renovación urbanas instrumentos de capital internacional, turismo global y migraciones de élites, cuyos centros históricos son los más afectados, debido a la competencia por el espacio urbano (Delgado et al., 2025). Si bien la forma metodológica para estudiar la gentrificación ha sido principalmente cualitativa, y más recientemente con un enfoque cuantitativo, resulta indispensable utilizar modelados predictivos para detectar

tendencias urbanas, basándose en datos censales (Maya et al., 2024; Yee & Dennett, 2022; Owens, 2012; Wei & Knox, 2013). Los modelos basados en Machine Learning tienen gran capacidad para manejar datos de alta dimensionalidad y encontrar patrones gracias a su capacidad de aprendizaje y mejoramiento a partir de los datos. Además, esta técnica permitirá identificar barrios o vecindarios expuestos a posibles procesos de gentrificación. Este estudio propone identificar los barrios gentrificados y propensos a la gentrificación (gentrificables) de Cartagena mediante un modelo de aprendizaje de ensamblaje (Bootstrap Aggregating) Random Forest, perteneciente al grupo de modelos de aprendizaje supervisado, donde el aprendizaje parte de los datos etiquetados —en este caso, la etiqueta “gentrificación”, construida a partir de criterios socioeconómicos con datos censales en los periodos de 2005 y 2018 (Loukaitou-Sideris et al., 2019).

2. CASOS DE ESTUDIO

En el estudio Understanding Urban Gentrification through Machine Learning: Predicting Neighbourhood Change in London (Reades et al., 2019), se realiza un análisis socioeconómico de los procesos de transición y los patrones de cambio en los barrios de Londres, utilizando datos censales de 2001 y 2011 para predecir las zonas con mayor probabilidad de gentrificación hacia 2021. Para ello, se aplica un modelo de Random Forest, destacando su baja complejidad en la hiperparametrización, su capacidad para reducir el sesgo y la forma en que la aleatorización de las muestras contribuye a evitar el sobreajuste.

Los resultados muestran una mejora sostenida en el Este Interior de Londres y su expansión hacia los distritos exteriores, mientras que algunas zonas periféricas presentan signos de deterioro. Los cambios en el estatus de los barrios se asocian principalmente con factores económicos y laborales, más que con aspectos del entorno físico. A pesar de la incertidumbre generada por factores políticos, como el Brexit, el estudio evidencia el potencial del aprendizaje automático para anticipar transformaciones urbanas y propone una integración entre enfoques cualitativos y cuantitativos con el fin de promover procesos de regeneración urbana que no impliquen desplazamiento social. Otro estudio relevante es Building a Predictive Machine Learning Model of Gentrification in Sydney (Thackway et al., 2023). En esta investigación se emplea un modelo de machine learning basado en árboles de decisión para predecir el cambio urbano en Sídney, utilizando diversos índices socioeconómicos. El modelo predice la gentrificación mediante una combinación de información censal y no censal correspondiente a los años 2011 y 2016, alcanzando un nivel de certeza del 74,7% según el indicador AUC-ROC. Posteriormente, se realiza una extrapolación hacia 2021, la cual evidencia una expansión desde el centro de la ciudad hacia las zonas periféricas, principalmente en Homebush, Bankstown, Auburn, Ryde y Sutherland. Para el modelado se emplearon datos sobre precios de vivienda y desarrollo urbano, los cuales confirman un efecto de desplazamiento que se propaga más allá del centro urbano. Finalmente, el estudio discute la necesidad de regular las políticas relacionadas con la vivienda pública, complementándolas con medidas de control de alquileres e impuestos redistributivos.

El estudio Identifying Gentrification using Machine Learning (Yoo & Census Bureau, 2023) explora técnicas de Machine learning para predecir unidades de viviendas propensas al riesgo de la gentrificación, usando encuestas de hogares “American Housing Survey” (AHS, por sus siglas en inglés) para el área metropolitana de de Wasgington D.C, cuyos datos fueron proporcionados por Metropolitan Statistical Area (MSA), American Community Survey, Commercial real estate website y de páginas web (riskfactor.com, climatecheck.com, Walkscore.com y GreatSchools.org). El AHS son datos de tipo panel que contiene información socioeconómica de las viviendas y sus características físicas que permiten evaluar que tan probable es que dichas viviendas sean desplazadas. Los periodos utilizados fueron 2015, 2017 y 2019, se usaron variables la edad, educación, estado civil, características habitacionales de las viviendas, servicios cercanos, riesgos ambientales, accesibilidad urbana e ingresos anteriores y actuales del hogar. El articulo construye la variable de “Gentrificado” usando 3 crítiterios principales: 1) Todos los miembros de los hogares en 2017 y 2019 son residentes distintos a los 2015, 2) el crecimiento del ingreso del hogar es mayor a la tasa de crecimiento a nivel censal, 3) La llegada de los nuevos

residentes es a causa de mejores empleos, viviendas o vecindarios. Se identificaron 250 unidades gentrificadas. Una vez se obtuvo la variable etiquetada Y (*Gentrificado*) se realizaron 6 modelos de clasificación para la identificación de la gentrificación previa: Logistic Regression (LR), K-nearest Neighbors Classifier (KNN), Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GB). Se resalta que los mejores modelos son el Random Forest y el Gradient Boosting, el primero posee un conjunto de árboles de decisiones que operan de manera paralela con una muestra aleatoria de los datos (bootstrap) y con un subconjunto aleatorio de variables.

Luego, el resultado final se obtiene por votación (clasificación) o promedio (regresión), que combina arboles de decisión minimizando el gradiente de error al corregir los errores de los árboles anteriores de manera secuencial que captura relaciones no lineales complejas, contiene alta precisión predictiva, y maneja la colinealidad y los valores atípicos. El resultado mostró que el performance del mejor modelo fue el de Random Forest, con un Accuracy de 0,83, una precisión de 0,81, recall de 0,87 y un F1 Score de 0,84. Finalmente se clasifico 3 categorías de genrtificación: alto riesgo, medio riesgo y bajo riesgo. El estudio destaca el uso del nowcasting o predicción en tiempo real de la gentrificación, mostrando que en el área metropolitana de Washington D.C. este proceso está impulsado principalmente por jóvenes adultos con alta formación académica que buscan apartamentos en zonas urbanas caminables. En conjunto, demuestra que aplicar modelos de inteligencia artificial para anticipar la gentrificación, en lugar de analizarla solo después de ocurrida, puede ayudar a los responsables de política pública a actuar de forma preventiva y diseñar estrategias más efectivas frente a este fenómeno urbano.

En el estudio Stratifying and predicting patterns of neighbourhood change and gentrification: An urban analytics approach (Yee & Dennett, 2022) se hace uso del machine learning aplicando el logatirmo de Random Forest para modelar los patrones de mejora o revalorización de los barrios de Londres, mediante datos censales, para luego predecir estados de vecindarios hacia el 2021, las bases de datos utilizadas son The Office for Natioal Statistics (ONS), The Consumer Data Research Centre (CDRC) y The Greater London Authority (GLA). Las variables utilizadas fueron la composición socioeconómica de los vecindarios, las características locales de la vivienda, nuevas construcciones residenciales, reconversiones o rehabilitaciones de viviendas existentes, registros electorales, bases de datos de consumo, transacciones inmobiliarias y rotaciones poblacionales, los periodos utilizados fueron 2001 y 2011. El estudio identificó y clasificó diversos tipos de cambios urbanos en Londres para los años anteriormente mencionados, se destaca que hubo un aumento de la gentrificación en el centro de Londres, principalmente en los distritos de Hammersmith, Fulham, Kensington y Chelsea y Newham, sin embargo, también se presentó el fenómeno de pérdida de cohesión social y desplazamientos de grupos de bajo ingresos por lo de alto ingresos, finalmente el artículo propone considerar políticas urbanas más equitativas.

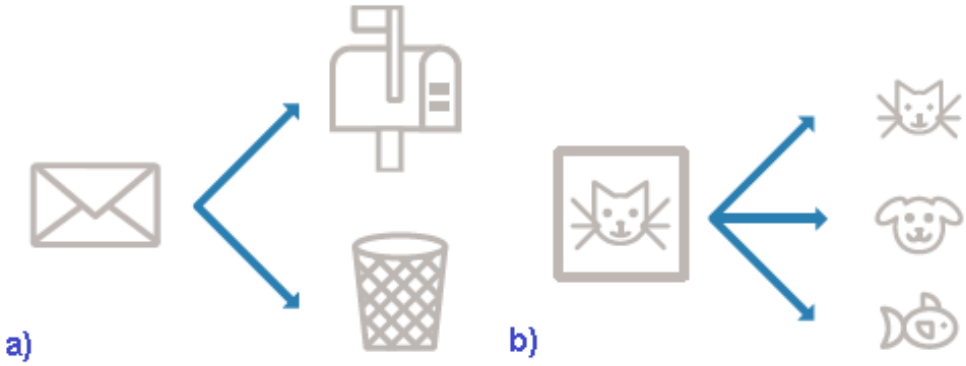
3. MARCO TEORICO Y REFERENCIAL

En esta sección se presentan los fundamentos teóricos esenciales que sustentan el modelo de clasificación propuesto para la predicción de gentrificación. Se abordan de forma concisa los conceptos fundamentales del aprendizaje automático y los modelos de clasificación, estableciendo las diferencias entre aprendizaje supervisado y no supervisado junto con sus respectivas ventajas y limitaciones. Posteriormente, se realiza una revisión de los principales algoritmos de clasificación existentes en la literatura, para finalmente profundizar en los fundamentos teóricos del algoritmo Random Forest, que constituye el modelo empleado en la presente investigación, detallando su estructura, funcionamiento y propiedades que lo hacen idóneo para problemas de clasificación complejos.

3.1 Modelos de clasificación con Machine Learning

El aprendizaje automático o machine learning (ML) es un campo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos capaces de aprender patrones a partir de datos de entrada, sin ser explícitamente programados para cada tarea específica. Según Shalev-Shwartz & Ben-David (2014), el aprendizaje puede entenderse como el proceso de convertir experiencia en conocimiento, donde la entrada de un algoritmo de aprendizaje son datos de entrenamiento que representan experiencia, y la salida es un programa que puede realizar alguna tarea. Los algoritmos de ML mejoran su desempeño en tareas específicas a medida que adquieren más experiencia, definiendo el "aspecto de aprendizaje" como el hecho de que mientras mejor se desempeña un algoritmo en una tarea específica, mejor ha aprendido de esa experiencia (Casali et al., 2022; Thackway et al., 2023). Los modelos de clasificación constituyen una categoría fundamental dentro del aprendizaje automático supervisado, cuyo objetivo es asignar instancias u observaciones a categorías o clases predefinidas (The MathWorks, 2016). La **Figura 1** muestra los dos tipos de problemas de clasificación abordados por estas técnicas.

Figura 1. a) Problema de clasificación binario. **b)** Problema de clasificación multiclase



Fuente: The MathWorks (2016).

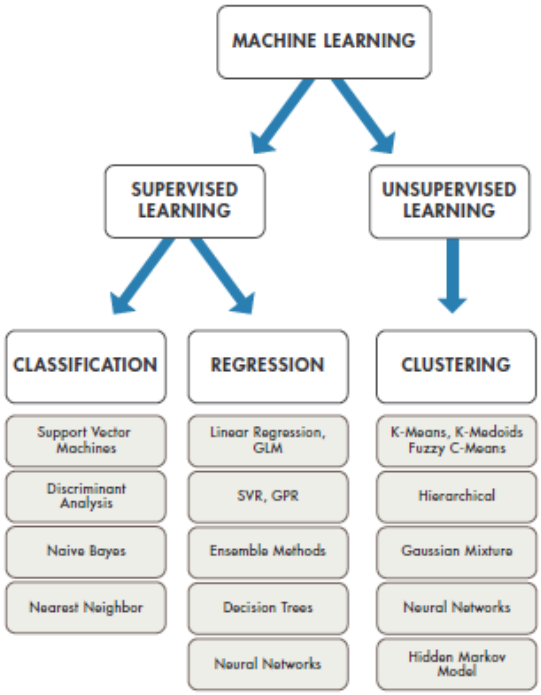
Formalmente, un problema de clasificación busca aprender una función predictora h: X → Y, donde X representa el espacio de características de entrada (conjunto de instancias) de Y corresponde al conjunto finito de etiquetas o clases posibles (Fernández-Delgado et al., 2014; Shalev-Shwartz & Ben-David, 2014).

En el caso más simple, la clasificación binaria involucra dos clases, típicamente representadas como Y = {0,1} o Y = {-1, +1}, sin embargo, muchos problemas del mundo real requieren clasificación multiclase, donde Y puede contener múltiples categorías; por ejemplo, en la clasificación de documentos según tema, X sería el conjunto de todos los documentos posibles e Y el conjunto de tópicos disponibles (Shalev-Shwartz & Ben-David, 2014; The MathWorks, 2016). El proceso de aprendizaje en modelos de clasificación se fundamenta en un conjunto de datos de entrenamiento S = {(x_1, y_1), ..., (x_m, y_m)}, donde cada par (x_i, y_i) consiste en una instancia x_i del espacio de características y su etiqueta correspondiente y_i (Fernández-Delgado et al., 2014; Shalev-Shwartz & Ben-David, 2014). A partir de estos ejemplos etiquetados, el algoritmo de aprendizaje debe generar un clasificador o hipótesis h que pueda predecir con precisión las etiquetas de nuevas instancias no observadas durante el entrenamiento, este clasificador representa la regla de decisión aprendida que el modelo utilizará para clasificar datos futuros (Fernández-Delgado et al., 2014). La capacidad de generalización (es decir, la habilidad del modelo para realizar predicciones correctas sobre datos nuevos) constituye un aspecto crucial que distingue el aprendizaje efectivo de la simple memorización de los datos de entrenamiento (Fernández-Delgado et al., 2014; Shalev-Shwartz & Ben-David, 2014; The MathWorks, 2016).

3.1.1 Aprendizaje supervisado y no supervisado

Los algoritmos de aprendizaje automático se dividen en dos grandes grupos según la naturaleza de los datos de entrenamiento y el tipo de interacción entre el aprendiz y el entorno: aprendizaje supervisado y aprendizaje no supervisado (ver **Figura 2**) (Casali et al., 2022; Shalev-Shwartz & Ben-David, 2014; The MathWorks, 2016).

Figura 2. Técnicas de machine learning



Fuente: (The MathWorks, 2016).

3.1.1.1 Aprendizaje supervisado

El aprendizaje supervisado describe un escenario en el cual los ejemplos de entrenamiento contienen información adicional significativa (las etiquetas o labels) que está ausente en los ejemplos de prueba a los que se aplicará el modelo aprendido (Fernández-Delgado et al., 2014; Shalev-Shwartz & Ben-David, 2014). En este enfoque, el algoritmo utiliza un conjunto de entrenamiento de ejemplos con respuestas correctas predefinidas, donde cada instancia x_i se presenta junto con su etiqueta correspondiente y_i, permitiendo al modelo aprender la relación entre las características de entrada y las salidas deseadas (Casali et al., 2022). En este contexto, puede conceptualizarse el entorno como un "profesor" que supervisa al aprendiz proporcionando la información adicional necesaria (Shalev-Shwartz & Ben-David, 2014).

Como ejemplo ilustrativo, considérese la tarea de detección de correos spam: el algoritmo recibe correos electrónicos de entrenamiento para los cuales se proporciona la etiqueta spam/no-spam, y con base en este entrenamiento, debe inferir una regla para etiquetar nuevos mensajes entrantes (Shalev-Shwartz & Ben-David, 2014). La experiencia adquirida tiene como objetivo predecir la información faltante (las etiquetas) para los datos de prueba, permitiendo al modelo clasificar instancias no vistas durante el entrenamiento. Maya et al. (2024) señalan que, en modelos supervisados, el entrenamiento se basa en ejemplos presentados por el usuario (datos etiquetados), lo que constituye una diferencia fundamental con respecto a enfoques no supervisados.

3.1.1.2 Ventajas y limitaciones del aprendizaje supervisado

El aprendizaje supervisado presenta como principal ventaja un objetivo claramente definido: aprender un clasificador que prediga las etiquetas de ejemplos futuros con la mayor precisión posible (Shalev-Shwartz & Ben-David, 2014). Además, un aprendiz supervisado puede estimar el éxito o riesgo de sus hipótesis utilizando los datos de entrenamiento etiquetados mediante el cálculo de la pérdida empírica, proporcionando un mecanismo directo de evaluación del desempeño del modelo (Shalev-Shwartz & Ben-David, 2014; The MathWorks, 2016). Sin embargo, la principal limitación del aprendizaje supervisado radica en la necesidad de contar con datos etiquetados para el entrenamiento. La obtención de estas etiquetas frecuentemente requiere intervención humana experta, lo cual puede resultar costoso, consumir tiempo considerable, y en algunos casos ser prácticamente inviable para conjuntos de datos de gran escala (Maya et al., 2024). Esta dependencia de datos etiquetados representa una restricción significativa que puede limitar la aplicabilidad del enfoque supervisado en ciertos contextos.

3.1.1.3 Aprendizaje no supervisado

En contraste, el aprendizaje no supervisado opera sobre datos sin etiquetas, donde las respuestas correctas no están disponibles durante el proceso de entrenamiento (Casali et al., 2022). En este paradigma, no existe distinción entre datos de entrenamiento y datos de prueba; en su lugar, el aprendiz procesa datos de entrada con el objetivo de generar algún tipo de resumen o versión comprimida de esos datos (Shalev-Shwartz & Ben-David, 2014). Maya et al. (2024) explican que los modelos no supervisados aprenden de los patrones intrínsecos de datos no

etiquetados, utilizando la estructura misma de los datos para generar predicciones. En lugar de predecir etiquetas específicas, el objetivo es organizar los datos de manera significativa, siendo el clustering o agrupamiento una tarea típica de este enfoque (Shalev-Shwartz & Ben-David, 2014) Como ejemplo, en la tarea de detección de anomalías, el algoritmo recibe únicamente un gran volumen de mensajes de correo electrónico sin etiquetas, y su tarea consiste en detectar mensajes "inusuales" basándose en los patrones encontrados en los datos (Shalev-Shwartz & Ben-David, 2014). Los enfoques no supervisados se emplean comúnmente para clustering, reducción de dimensionalidad y estimación de densidad (Maya et al., 2024).

3.1.1.4 Ventajas y limitaciones del aprendizaje no supervisado

La ventaja fundamental del aprendizaje no supervisado es que no requiere datos etiquetados, eliminando así la necesidad de un proceso costoso de etiquetado manual (Shalev-Shwartz & Ben-David, 2014). Esto lo hace particularmente útil para tareas de exploración de datos donde el objetivo es descubrir estructura o patrones inherentes sin conocimiento previo de las categorías existentes.

No obstante, el aprendizaje no supervisado enfrenta desafíos importantes. La ausencia de "verdad absoluta" (ground truth) constituye un problema común: no existen etiquetas que predecir, y consecuentemente, no hay un procedimiento claro de evaluación del éxito del algoritmo (Shalev-Shwartz & Ben-David, 2014). Incluso con conocimiento completo de la distribución subyacente de los datos, no resulta evidente cuál sería el clustering "correcto" o cómo evaluar un agrupamiento propuesto (Shalev-Shwartz & Ben-David, 2014). Esta ambigüedad en la definición de éxito representa una limitación fundamental que dificulta tanto el desarrollo como la evaluación de modelos no supervisados, ya que un mismo conjunto de datos puede admitir múltiples soluciones de clustering igualmente válidas, pero conceptualmente diferentes (Shalev-Shwartz & Ben-David, 2014).

3.1.2 Principales algoritmos de clasificación

Existe una amplia diversidad de algoritmos de clasificación disponibles en la literatura. Fernández-Delgado et al. (2014) realizaron una evaluación exhaustiva de 179 clasificadores provenientes de 17 familias diferentes, utilizando 121 conjuntos de datos de la base UCI. Las familias evaluadas incluyen: análisis discriminante, métodos bayesianos, redes neuronales, máquinas de vectores de soporte (SVM), árboles de decisión, clasificadores basados en reglas, boosting, bagging, stacking, random forests y otros ensambles, modelos lineales generalizados, vecinos más cercanos, regresión de mínimos cuadrados parciales y regresión de componentes principales, regresión logística y multinomial, y splines de regresión adaptativa múltiple (MARS).

En aplicaciones prácticas, Casali et al. (2022) identificaron que los algoritmos supervisados más frecuentemente utilizados en análisis urbanos son: redes neuronales (NN), Random Forests (RF), máquinas de vectores de soporte (SVM), árboles de decisión con gradient boosting (GBDT), árboles de decisión (DT), K-vecinos más cercanos (KNN) y regresión logística. Entre estos, los resultados de Fernández-Delgado et al. (2014) demuestran que Random Forest es la mejor familia de clasificadores, con 3 de los 5 mejores clasificadores pertenecientes a esta familia. El mejor clasificador Random Forest alcanza 94.1% de la precisión máxima, superando el 90% en el 84.3% de los conjuntos de datos evaluados. Le

sigue el SVM con kernel gaussiano implementado en LibSVM con 92.3% de la precisión máxima, aunque la diferencia no es estadísticamente significativa. Otros modelos destacados incluyen SVM con kernels gaussiano y polinomial, extreme learning machine con kernel gaussiano, C5.0 y avNNet (comité de perceptrones multicapa). La familia SVM posiciona 4 clasificadores en el top-10, mientras que redes neuronales y ensambles de boosting colocan 5 y 3 miembros respectivamente en el top-20 (Fernández-Delgado et al., 2014)

3.1.3 Random Forest

Random Forest (RF) es un algoritmo de aprendizaje por ensamble (ensemble learning) introducido por Breiman (2001) que ha alcanzado reconocimiento como uno de los métodos de clasificación y regresión más efectivos en el aprendizaje automático supervisado (Fernández-Delgado et al., 2014). Este algoritmo constituye una extensión del método de bagging (bootstrap aggregating) aplicado a árboles de decisión (Breiman, 2001), donde múltiples árboles individuales son entrenados de manera independiente sobre submuestras bootstrap del conjunto de datos original, y sus predicciones son posteriormente agregadas mediante votación mayoritaria en clasificación o promediación en regresión (Mansour & Schain, 2001). La principal innovación de Random Forest radica en la introducción de una fuente adicional de aleatoriedad durante la construcción de cada árbol: en cada nodo del árbol, en lugar de considerar todas las variables predictoras disponibles para determinar la mejor división, el algoritmo selecciona aleatoriamente un subconjunto de variables candidatas (Breiman, 2001), lo que incrementa la diversidad entre los árboles del ensamble y reduce la correlación entre sus predicciones (Shalev-Shwartz & Ben-David, 2014).

3.1.3.1 Fundamentos: Árboles de Decisión como Clasificadores Base

Los árboles de decisión constituyen el componente fundamental sobre el cual se construye Random Forest (Shalev-Shwartz & Ben-David, 2014). Un árbol de decisión es un modelo de predicción representado por una estructura de árbol (ver figura 3) donde cada nodo interno corresponde a una prueba sobre una variable de entrada, cada rama representa el resultado de dicha prueba, y cada nodo hoja (terminal) contiene una etiqueta de clase o un valor de predicción (Shalev-Shwartz & Ben-David, 2014). Formalmente, un árbol de decisión implementa una función predictora h que particiona recursivamente el espacio de características X mediante una serie de reglas de decisión binarias. Para variables continuas y categóricas, estas reglas tienen la forma:

Regla de decisión para variables continuas:

1 si x_i < θ

0 en caso contrario

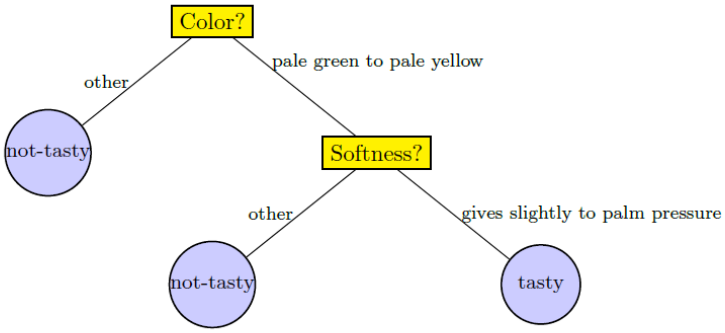
Regla de decisión para variables categóricas:

1 si x_i = v

0 en caso contrario

donde x_i representa la i-ésima característica y θ o v son umbrales o valores específicos determinados durante el entrenamiento (Shalev-Shwartz & Ben-David, 2014).

Figura 3. Esquema árbol de decisión



Fuente: (Shalev-Shwartz & Ben-David, 2014).

La construcción de un árbol de decisión sigue un enfoque voraz (greedy) donde, partiendo de un nodo raíz que contiene todo el conjunto de entrenamiento, se selecciona iterativamente la variable y el punto de división que optimizan localmente algún criterio de calidad, subdividiendo progresivamente los datos hasta alcanzar un criterio de parada (Mansour & Schain, 2001). Entre los algoritmos más conocidos para el crecimiento de árboles se encuentran ID3 (Iterative Dichotomizer 3) desarrollado por Quinlan (1993) y CART (Classification and Regression Trees) propuesto por Breiman (2001). El algoritmo ID3 construye el árbol mediante llamadas recursivas donde en cada nodo se calcula una medida de ganancia (Gain) para todas las variables disponibles, seleccionando aquella que maximiza dicha ganancia (Quinlan, 1993). Por su parte, el algoritmo CART utiliza divisiones binarias y emplea criterios como el índice de Gini para problemas de clasificación (Breiman, 2001).

3.1.3.2 Criterios de División en Árboles de Decisión

La selección de la variable y punto de división óptimos en cada nodo del árbol se fundamenta en la evaluación de diferentes medidas de ganancia, las cuales cuantifican la reducción en impureza o incertidumbre lograda al particionar el conjunto de datos (Probst et al., 2019). Las tres medidas más utilizadas en la literatura son el error de entrenamiento, la ganancia de información basada en entropía, y el índice de Gini (Mansour & Schain, 2001; Shalev-Shwartz & Ben-David, 2014). Para un conjunto de entrenamiento S y una variable predictora i, la ganancia se define como la diferencia entre la impureza antes y después de la división.

3.1.3.2.1 Criterio 1: Error de Entrenamiento

La medida más simple es la reducción en error de entrenamiento, donde la función de costo se define como (Chen & Guestrin, 2016; Shalev-Shwartz & Ben-David, 2014):

$$C(a) = min\{a, 1-a\}$$

El error de entrenamiento antes de dividir según la variable i es $C(PS[y=1])$, y después de la división es:

$$PS[xi=1] \cdot C(PS[y=1|xi=1]) + PS[xi=0] \cdot C(PS[y=1|xi=0])$$

donde $PS[\cdot]$ denota la probabilidad empírica sobre S. Por consiguiente, la ganancia basada en error de entrenamiento se expresa como:

$$Gain(S,i) = C(PS[y=1]) - [PS[xi=1] \cdot C(PS[y=1|xi=1]) + PS[xi=0] \cdot C(PS[y=1|xi=0])]$$

3.1.3.2.2 Criterio 2: Ganancia de Información (Information Gain)

La ganancia de información, utilizada en los algoritmos ID3 y C4.5 de Quinlan (1993), se basa en la diferencia entre la entropía de las etiquetas antes y después de la división, reemplazando $C(a)$ en la expresión anterior por la función de entropía (Shalev-Shwartz & Ben-David, 2014):

$$C(a) = -a \cdot log(a) - (1-a) \cdot log(1-a)$$

Esta medida cuantifica la reducción en la incertidumbre sobre la clase de las instancias resultante de conocer el valor de la variable predictora i (Quinlan, 1993). La función de entropía alcanza su máximo cuando $a = 0.5$ (máxima incertidumbre) y su mínimo cuando $a = 0$ o $a = 1$ (certeza completa).

3.1.3.2.3 Criterio 3: Índice de Gini

El índice de Gini, empleado por el algoritmo CART de (Breiman, 2001), define la función de costo como:

$$C(a) = 2a(1-a)$$

Tanto la ganancia de información como el índice de Gini son cotas superiores suaves y cóncavas del error de entrenamiento (Shalev-Shwartz & Ben-David, 2014), propiedades que resultan ventajosas en diversas situaciones, particularmente en contextos donde la diferenciabilidad y la convexidad facilitan la optimización (Mansour & Schain, 2001). No obstante, Strobl et al. (2008) han demostrado que el índice de Gini presenta un sesgo hacia variables con mayor número de categorías o escalas de medición continuas.

3.1.3.3 Construcción del Ensemble mediante Bagging y Selección Aleatoria de Variables

Random Forest construye un ensamble de árboles de decisión mediante un procedimiento que combina dos fuentes principales de aleatoriedad: el muestreo bootstrap de las instancias de entrenamiento y la selección aleatoria de subconjuntos de variables en cada división (Breiman, 2001; Chen & Guestrin, 2016). Formalmente, dado un conjunto de entrenamiento:

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

Donde:

- $X_i \in \mathbb{R}^d$ representa el vector de características de la instancia i
- Y_i representa la etiqueta correspondiente
- m es el tamaño del conjunto de entrenamiento
- d es el número de variables predictoras

Random Forest genera T árboles de decisión $\{h_1, h_2, \ldots, h_T\}$ (Shalev-Shwartz & Ben-David, 2014), donde cada árbol h_T se construye mediante el siguiente procedimiento de dos pasos:

3.1.3.3.1 Paso 1: Muestreo Bootstrap

Se genera una submuestra bootstrap $S't$ de tamaño m' mediante muestreo con reemplazo de S usando la distribución uniforme (Breiman, 2001):

$$S't = muestra\ de\ tamaño\ m'\ extraída\ con\ reemplazo\ de\ S$$

Esto significa que algunas instancias originales pueden aparecer múltiples veces en $S't$ mientras otras pueden no aparecer en absoluto. Las instancias no seleccionadas en este proceso, denominadas observaciones out-of-bag (OOB), constituyen aproximadamente el 36.8% de los datos originales y desempeñan un papel importante en la evaluación del modelo sin requerir un conjunto de validación adicional (Breiman, 2001).

3.1.3.3.2 Paso 2: Selección Aleatoria de Variables en Cada Nodo

Durante el crecimiento de cada árbol sobre la muestra bootstrap $S't$, en cada etapa de división se introduce una segunda fuente de aleatoriedad (Breiman, 2001): en lugar de evaluar todas las d variables disponibles para determinar la mejor división, el algoritmo selecciona aleatoriamente un subconjunto:

$$It \subseteq \{1, 2, \ldots, d\}$$

de tamaño k (comúnmente denominado *mtry* en las implementaciones), y la variable de división se elige únicamente de entre este subconjunto restringido maximizando la ganancia:

$$variable\ óptima = argmax\ Gain(S't, i)\ para\ i \in It$$

Este procedimiento se repite de manera independiente en cada nodo del árbol, generando una nueva selección aleatoria de variables candidatas (Breiman, 2001). Estas dos fuentes de aleatoriedad (muestreo bootstrap y selección aleatoria de variables) trabajan conjuntamente para reducir la correlación entre los árboles individuales del ensamble (Strobl et al., 2008). Intuitivamente, si k es pequeño en comparación con d, esta restricción puede prevenir el sobreajuste al limitar la capacidad de cada árbol individual para memorizar completamente los datos de entrenamiento (Shalev-Shwartz & Ben-David, 2014)

La fundamentación teórica de por qué Random Forest alcanza un desempeño superior al de los árboles individuales se encuentra en los principios del bagging y en la teoría de métodos de ensamble. El método bagging (bootstrap aggregating), introducido por Breiman (2001), aprovecha el hecho de que los árboles de decisión son clasificadores inestables pero que, en promedio, producen predicciones correctas. Mediante la combinación de predicciones de un conjunto diverso de árboles, bagging utiliza esta inestabilidad para reducir la varianza de la predicción sin incrementar sustancialmente el sesgo (Breiman, 2001). Los resultados teóricos de Peter Bühlmann (2002) demostraron que la mejora en la precisión de predicción de los ensambles se logra mediante el suavizado (smoothing) de las fronteras de decisión rígidas creadas por las divisiones en árboles de clasificación individuales, lo cual reduce efectivamente la varianza de la predicción. En Random Forest, la introducción de la selección aleatoria de variables en cada división genera aún mayor diversidad entre los árboles (Strobl et al., 2008), permitiendo que variables predictoras que de otro modo serían eclipsadas por competidores más fuertes puedan entrar en el ensamble, revelando efectos de interacción que podrían haberse perdido (Strobl et al., 2008).

3.1.3.4 Mecanismo de Predicción

La predicción final de Random Forest para una nueva instancia x se obtiene mediante la agregación de las predicciones individuales de todos los árboles del ensamble (Breiman, 2001).

3.1.3.4.1 Para Problemas de Clasificación:

Se utiliza votación por mayoría: cada árbol ht(x) emite un voto para una clase específica, y la clase con el mayor número de votos es seleccionada como la predicción final (Svetnik et al., 2003). Formalmente:

$$h_{RF}(x) = argmax_{c \in Y} \sum_{t=1}^{T} 1[h_t(x) = c]$$

Donde Y representa el conjunto de posibles clases T es el número total de árboles en el bosque $1[ht(x) = c]$ es la función indicadora que vale 1 si el árbol t predice la clase c, y 0 en caso contrario.

3.1.3.4.2 Para Problemas de Regresión:

La predicción se obtiene promediando las predicciones numéricas de todos los árboles (Krstajic et al., 2014):

$$h_{RF}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

donde $h_t(x)$ es el valor numérico predicho por el árbol t para la instancia x. Este proceso de agregación constituye la esencia del método de ensamble (Chen & Guestrin, 2016), permitiendo que Random Forest aproveche la sabiduría colectiva de múltiples árboles para producir predicciones más robustas y precisas que cualquier árbol individual (Breiman, 2001).

3.1.3.5 Hiperparámetros y su Influencia en el Desempeño

El rendimiento de Random Forest está influenciado por diversos hiperparámetros que controlan tanto la estructura de los árboles individuales como la configuración global del ensamble (Probst et al., 2019). Los hiperparámetros principales incluyen:

(1) mtry (k): El número de variables candidatas consideradas en cada división, cuyo valor óptimo depende críticamente del número de variables relevantes para el problema (Probst et al., 2019).
(2) T (número de árboles): El número de árboles en el bosque, donde valores más altos generalmente conducen a mejor desempeño y mayor estabilidad en las medidas de importancia de variables, aunque con rendimientos decrecientes más allá de cierto punto (Breiman, 2001; Probst et al., 2019).
(3) sample_size: El tamaño de muestra para cada árbol, típicamente igual al tamaño del conjunto de entrenamiento n, aunque puede modificarse (Probst et al., 2019).
(4) replacement: El esquema de muestreo, que especifica si las muestras bootstrap se extraen con o sin reemplazo (Breiman, 2001).
(5) min_samples_split y min_samples_leaf: El tamaño mínimo de nodo, que controla cuándo se detiene la subdivisión de un nodo.
(6) max_depth: La profundidad máxima del árbol, que limita el crecimiento vertical del árbol.
(7) splitting_rule: El criterio de división, que puede ser Gini (Breiman, 2001), entropía (Quinlan, 1993), o variantes condicionales.

El hiperparámetro mtry es particularmente crítico, ya que controla directamente el trade-off entre la fortaleza individual de los árboles y su correlación mutua (Probst et al., 2019). Según (Probst et al., 2019), "la aleatoriedad utilizada en la construcción de árboles debe apuntar a baja correlación ρ mientras mantiene fortaleza razonable". Valores bajos de mtry producen árboles más diversos y menos correlacionados, mejorando la estabilidad al agregar (Probst et al., 2019), y también permiten explotar mejor variables con efectos moderados que serían enmascaradas por variables con efectos fuertes si estas últimas siempre fueran candidatas para división (Strobl et al., 2008). Sin embargo, valores muy bajos de mtry también pueden resultar en árboles individualmente débiles (Probst et al., 2019). Los valores típicos por defecto son:

Para clasificación: Para regresión:

$$mtry = \sqrt{p} \qquad mtry = p/3$$

donde p es el número total de variables predictoras (Breiman, 2001), aunque estos valores deben considerarse como puntos de partida que pueden requerir ajuste según las características específicas de los datos (Probst et al., 2019). El ajuste de hiperparámetros (hyperparameter tuning) puede realizarse mediante estrategias como búsqueda en rejilla (grid search), búsqueda aleatoria (random search), o métodos más sofisticados como optimización bayesiana (Jun, 2021), utilizando las observaciones out-of-bag para evaluar el desempeño sin necesidad de un conjunto de validación separado (Mansour & Schain, 2001).

3.1.3.6 Medidas de Importancia de Variables

Random Forest proporciona medidas de importancia de variables que cuantifican la contribución relativa de cada variable predictora a la capacidad predictiva del modelo (Breiman, 2001; Strobl et al., 2008), siendo esta una de las características más valiosas del algoritmo para aplicaciones donde la interpretabilidad es importante. Existen dos medidas principales:

- **Importancia Basada en Permutación (Permutation Importance):** La importancia por permutación se calcula mediante el siguiente procedimiento (Breiman, 2001):

  Para cada árbol en el bosque y cada variable predictora j:

  - Se permutan aleatoriamente los valores de la variable j en las observaciones out-of-bag
  - Se calcula la diferencia en precisión de predicción antes y después de la permutación
  - Esta diferencia se promedia sobre todos los árboles

  Formalmente, la importancia por permutación de la variable j puede expresarse como:

$$ImportanciaPerm(j) = \frac{1}{T}\sum_{t=1}^{T}[AccuracyOOB(t) - AccuracyOOB, perm(t,j)]$$

  Donde:

  - T es el número de árboles
  - AccuracyOOB(t) es la precisión del árbol t evaluada en sus observaciones out-of-bag
  - AccuracyOOB, perm(t,j) es la precisión después de permutar la variable j

  Una disminución grande en precisión indica que la variable es importante para la predicción (Breiman, 2001; Strobl et al., 2008).

- **Importancia Basada en Gini (Mean Decrease Impurity):** La importancia basada en Gini se calcula acumulando, para cada variable, la reducción total en impureza de Gini lograda por todas las divisiones que utilizan esa variable a lo largo de todos los árboles del bosque (Breiman, 2001).

  Para un árbol individual T, la importancia de la variable j se define como:

$$I^2j(T) = \sum_{n=1}^{T} 1[variable\ en\ nodo\ n = j] \times \Delta Impureza(n)$$

  Donde:

  - La suma se realiza sobre todos los nodos internos del árbol T
  - ΔImpureza(n) representa la reducción en impureza de Gini lograda en el nodo n

  La importancia final para Random Forest se obtiene promediando sobre todos los árboles (Mansour & Schain, 2001):

$$I^2j = \frac{1}{M} \sum_{m=1}^{T} I^2j(Tm)$$

  Donde M es el número de árboles en el bosque.

  Los valores de importancia resultantes se estandarizan típicamente para que sumen 100%, permitiendo una interpretación relativa directa (Breiman, 2001). Sin embargo, la importancia basada en Gini presenta un sesgo conocido: tiende a favorecer variables con muchas categorías o con escalas de medición continuas sobre variables binarias (Strobl et al., 2008), y puede sobrestimar la importancia de variables correlacionadas (Strobl et al., 2008). Por esta razón, la importancia por permutación es generalmente preferida cuando se requiere una evaluación más fiable del impacto verdadero de cada variable, particularmente en presencia de variables predictoras correlacionadas (Strobl et al., 2008).

3.1.4 Ventajas y Limitaciones de Random Forest

Random Forest presenta numerosas ventajas que han contribuido a su amplia adopción en aplicaciones prácticas de aprendizaje automático (Fernández-Delgado et al., 2014). En primer lugar, el algoritmo es altamente efectivo para prevenir el sobreajuste: al obtener la predicción final agregada de múltiples árboles de decisión independientes, Random Forest reduce significativamente la varianza de un árbol individual, conduciendo a mejores predicciones sobre datos nuevos (Breiman, 2001). En segundo lugar, Random Forest exhibe notable flexibilidad, ya que no requiere preprocesamiento extensivo de datos como escalamiento de variables, transformaciones, o imputación elaborada de valores faltantes, y no asume ninguna forma funcional específica de la relación entre variables predictoras y respuesta, a diferencia de modelos paramétricos como la regresión lineal o logística (Jun, 2021).

En tercer lugar, el algoritmo es naturalmente robusto frente a variables irrelevantes y puede manejar efectivamente espacios de características de alta dimensionalidad (Breiman, 2001). En cuarto lugar, Random Forest puede capturar relaciones no lineales complejas e interacciones entre variables sin necesidad de especificarlas explícitamente en el modelo (Chen & Guestrin, 2016). En quinto lugar, las observaciones out-of-bag proporcionan una estimación interna del error de generalización sin requerir un conjunto de validación separado, lo cual es computacionalmente eficiente (Breiman, 2001). Finalmente, Fernández-Delgado et al. (2014) demostraron empíricamente, mediante una evaluación exhaustiva de 179 clasificadores de 17 familias diferentes sobre 121 conjuntos de datos de la base UCI, que Random Forest es la mejor familia de clasificadores, con el mejor Random Forest alcanzando 94.1% de la precisión máxima y superando el 90% de precisión en el 84.3% de los conjuntos de datos evaluados.

No obstante, Random Forest también presenta limitaciones importantes que deben considerarse. La principal desventaja es su complejidad computacional: el entrenamiento de múltiples árboles de decisión completos requiere recursos computacionales considerables, especialmente con conjuntos de datos grandes y números elevados de árboles, aunque este costo puede mitigarse parcialmente mediante implementaciones paralelas dado que los árboles se entrenan independientemente (Breiman, 2001).

Adicionalmente, la interpretabilidad del modelo se ve significativamente reducida en comparación con un árbol de decisión individual: mientras que un solo árbol proporciona una representación visual clara y transparente de las reglas de decisión (Krstajic et al., 2014), Random Forest constituye esencialmente una "caja negra" donde la lógica de predicción resulta difícil de explicar a usuarios no técnicos, aunque las medidas de importancia de variables ofrecen cierta información sobre qué factores contribuyen más a las predicciones (Breiman, 2001; Probst et al., 2019). Finalmente, en comparación con algunos métodos de boosting como Gradient Boosting Decision Trees, Random Forest puede alcanzar precisión predictiva ligeramente inferior en ciertos conjuntos de datos, particularmente aquellos donde las señales predictivas son débiles o donde existe alta complejidad en las interacciones entre variables (Krstajic et al., 2014; Lundberg & Lee, 2017).

4. METODOLOGÍA

En este apartado se describe el proceso de compilación, depuración y transformación de los datos utilizados en el análisis. Se detallan las variables seleccionadas para la construcción del modelo y los patrones urbanos observados entre los años 2005 y 2018. Asimismo, se explica la creación de la variable dependiente correspondiente al fenómeno de gentrificación, la cual constituye el eje central del modelo de Random Forest. Posteriormente, se presentan los criterios socioeconómicos empleados para clasificar los barrios o vecindarios según su condición de *gentrificados* o *no gentrificados*. Finalmente, se expone la estimación del modelo de Random Forest, junto con sus indicadores de rendimiento y precisión, que permiten evaluar la capacidad predictiva del modelo y la validez de los resultados obtenidos.

4.1 Recopilación de datos

Las bases de datos utilizadas para este estudio parten de dos fuentes principales: datos censados del Departamento Administrativo Nacional de Estadística (DANE) y la Secretaría de Hacienda Distrital, para los años comprendidos durante el 2005 y 2018. Utilizando información socioeconómica, sociodemográfica y valor promedio del suelo por metro cuadrado.

4.1.1 DANE

Variables sociodemográficas y socioeconómicas

- Años promedio de escolaridad
- Número de viviendas ocupadas
- Índice de envejecimiento
- Población total

4.1.2 Secretaría de Hacienda Distrital

- Precio por m2

Los datos son extraídos por medio de REDATAM (Recuperación de Datos para Áreas Pequeñas), sistema de difusión de datos censales y estadísticos desarrollados por la CEPAL, el cual utiliza el DANE para realizar consultas y análisis de datos de gran volumen, y la información de valor del suelo por m2 fue proporcionada por la Secretaría de Hacienda Distrital, ambas a nivel manzana, que luego fueron colapsadas a nivel de los barrios de Cartagena.

4.2 Estructuración y transformación de datos

Con el objetivo de identificar los cambios urbanos y los patrones de transformación en los barrios, se calcularon primero las proporciones de viviendas ocupadas en relación con la población total de cada periodo analizado. Posteriormente, se estimó el índice de envejecimiento, definido como la razón entre la población de personas mayores de 60 años y la población joven (de 0 a 14 años), multiplicada por 100.

Un valor elevado de este índice indica una población envejecida, mientras que valores bajos reflejan una estructura poblacional más joven. Por su parte, el promedio de años escolares y el precio promedio por metro cuadrado se mantuvieron en términos absolutos, al considerarse variables de referencia estructural. Finalmente, se calcularon las variaciones intertemporales de cada indicador con el fin de construir una base de datos que permitiera evaluar la dirección y magnitud de los cambios ocurridos en los barrios a lo largo del periodo de estudio.

4.3 Creación del indicador de gentrificación

Se utilizaron los siguientes criterios para identificar si el área censada había experimentado un proceso de gentrificación:

a) La variación de la proporción de la vivienda ocupada es mayor al promedio que el de la ciudad
b) La variación de años promedios de escolaridad es mayor al promedio que el de la ciudad
c) La variación del precio del suelo por m2 es mayor al promedio que el de la ciudad
d) El índice de envejecimiento es menor que el percentil 25 (cuartil inferior) que el índice de envejecimiento de la ciudad

Tomando como referencia los criterios propuestos por Chapple et al. (2017) para identificar los vecindarios *gentrificados o en proceso de gentrificación*, este estudio adapta dicha metodología al contexto urbano de Cartagena, Colombia. En la **Tabla 1** se presentan los criterios discutidos.

Tabla 1. Criterios para clasificar barrios como “vulnerables” o “gentrificados/en proceso de gentrificación”

Clasificación	Indicadores
Vecindario vulnerable a la gentrificación (Año base) – Cumple a menos 3 indicadores	
% de hogares de bajos ingresos (menos del 80% del ingreso mediano del condado)	Por encima de la mediana del condado
% con educación universitaria o superior	Por debajo del percentil 40 del condado
% de arrendatarios	Por encima de la mediana del condado
% de población blanca no hispana	Por debajo de la mediana del condado
Barrio gentrificado o en proceso de gentrificación (cambio Año base–Año final)	
% con educación universitaria o superior	Por encima del promedio del condado
Ingreso mediano del hogar	Por encima del promedio del condado
% de población blanca no hispana	Por encima del promedio del condado
Alquiler bruto mediano	Por encima del promedio del condado

Fuente: Adaptado de Loukaitou-Sideris et al. (2019).

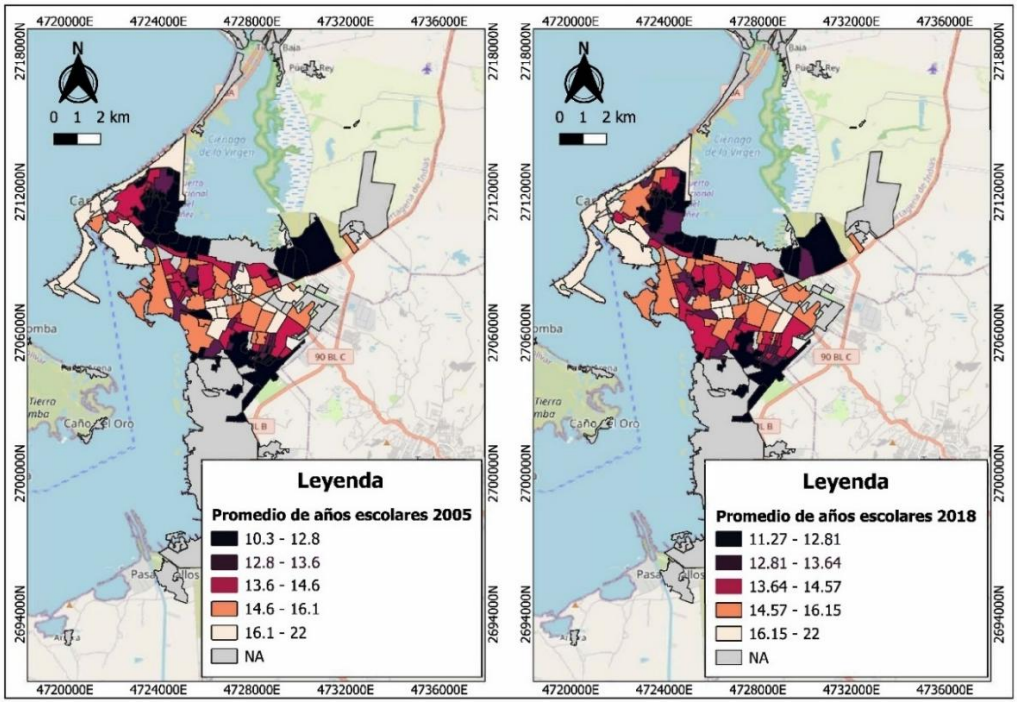
Si bien el modelo original fue diseñado para la ciudad de Los Ángeles, Estados Unidos (donde se dispone de variables como la proporción de población blanca no hispana y el ingreso de los hogares), la realidad urbana de Cartagena presenta dinámicas distintas, marcadas por procesos de turistificación y la llegada de nuevos inversores inmobiliarios que inciden directamente en la especulación del suelo.

Dadas las limitaciones de la información censal en Colombia, particularmente la ausencia de datos sobre ingresos de los hogares, la cual no se incluye en el modelo, además, se reemplazó la variable de educación superior utilizada en el modelo original por el índice de envejecimiento, considerando que en muchos barrios el desplazamiento no está asociado a un mayor acceso educativo, sino a transformaciones derivadas del mercado inmobiliario y turístico. En consecuencia, el modelo se ajustó empleando las variables disponibles para el contexto local, de modo que un barrio se clasificó como gentrificado cuando presentaba cambios significativos en al menos dos de los criterios seleccionados.

4.4 Perfil de los barrios gentrificados

Esta sección examina los patrones de gentrificación en la ciudad de Cartagena, mediante las variables seleccionadas y el cambio experimentado durante el periodo de 2005-2018, mediante mapas (ver **Figura 4**, **5, 6 y 7**) y así destacar los barrios con mayores incidencias en el marco del cambio urbano.

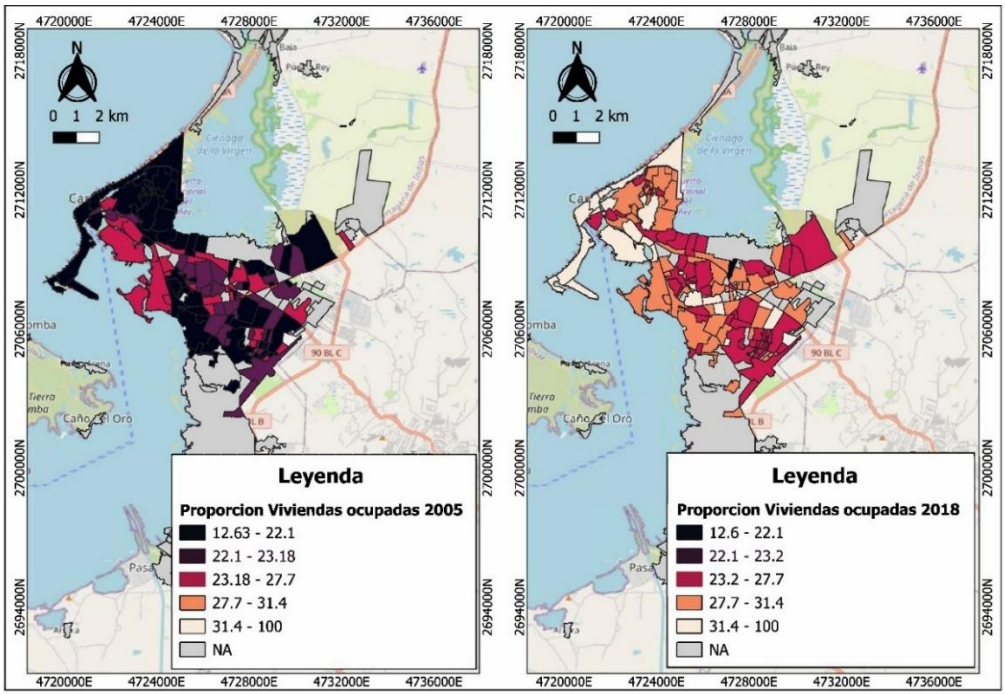
Figura 4. Cambio en el promedio de año escolar



Fuente: elaboración de los autores con base en DANE.

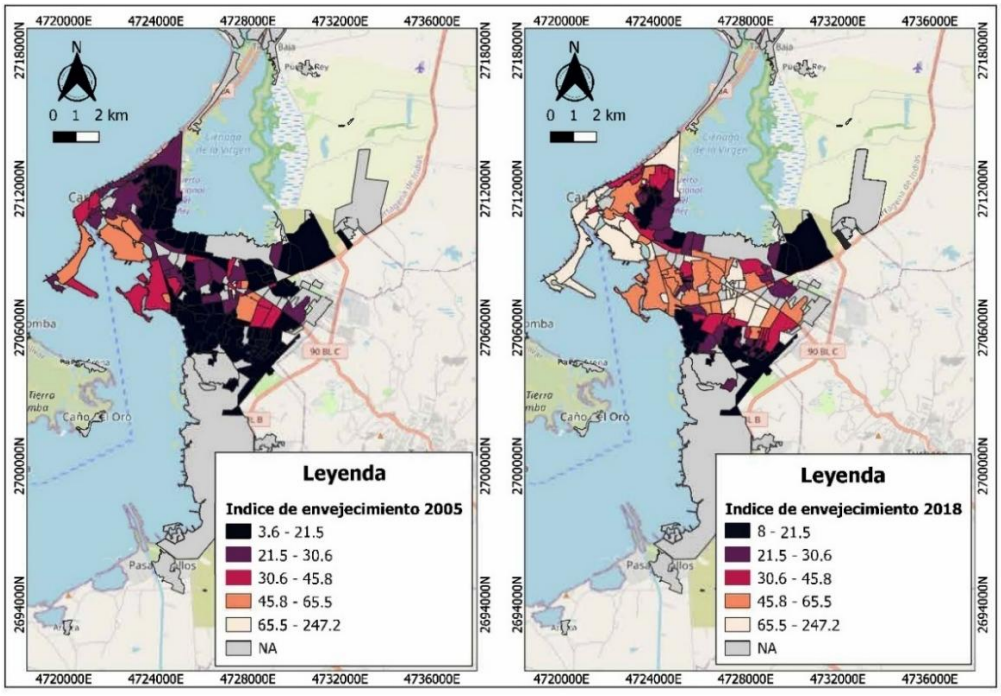
En 2005, los barrios con mayor promedio de años de escolaridad fueron La Matuna, Santa Mónica, Manga, San Pedro y El Cabrero, con cerca de 18 años en promedio. Para 2018, el panorama cambió, destacándose Chambacú (22 años), El Cabrero (20), Castillo Grande (19), Marbella (19) y Bocagrande (19) como los sectores con niveles educativos más altos. Los mayores incrementos en escolaridad se registraron en Chambacú (24%), seguido de El Laguito (15%), Castillo Grande (15%), Villa Hermosa (14%) y El Cabrero (13%) (ver **Figura 4**). En total, alrededor del 50% de los barrios (73) mostraron un aumento superior al promedio general, evidenciando un avance significativo en el nivel educativo de la población en buena parte del territorio.

Figura 5. Cambio en Proporción de viviendas ocupadas



Fuente: elaboración de los autores con base en DANE.

Figura 6. Cambio en el Índice de envejecimiento

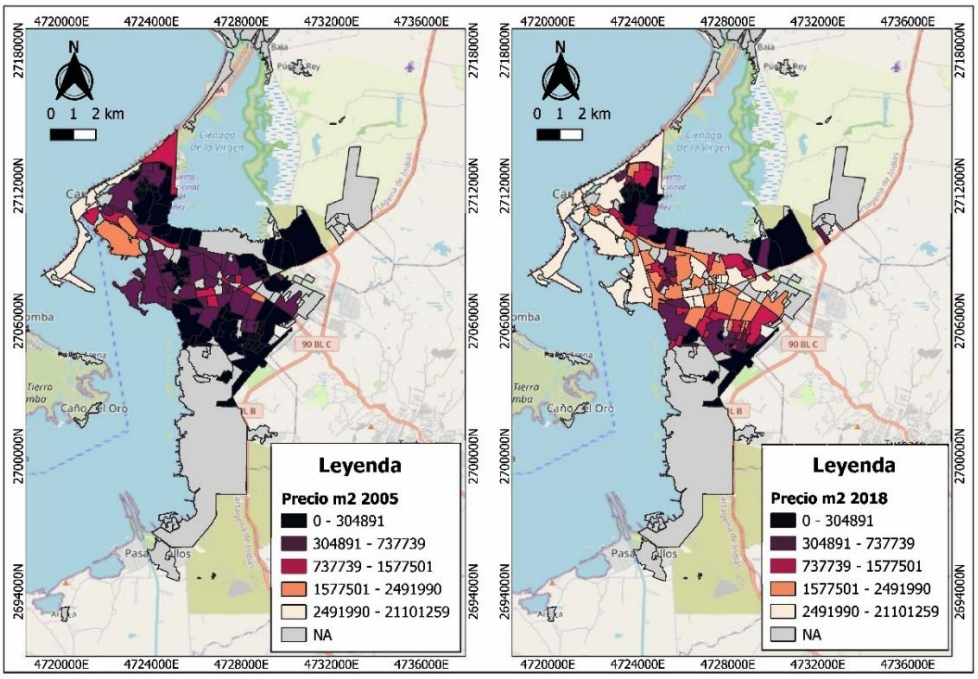


Fuente: elaboración de los autores con base en DANE.

En 2005, los cinco barrios con mayor proporción de viviendas ocupadas (ver **Figura 5**) eran Rubí (29%), El Laguito (28%), Armenia (27%), Chambacú (26%) y Pie del Cerro (26%). Para 2018, siendo el vecindario con mayor ocupación de viviendas Chambacú (100%), Centro (48%), La Matuna (48%), El Cabrero (43%) y El Laguito (43%). Se presentó mayor crecimiento en Chambacú, Centro, Santa María, Bocagrande y El Cabrero. El 29% (43) de los barrios obtuvo una variación mayor al promedio de la muestra. Se observa una mayor proporción de viviendas ocupadas en general, y principalmente en la zona norte y en el centro de la ciudad. Para el caso del índice de envejecimiento (ver **Figura 6**) algunos de los barrios se mantuvieron en los mismos umbrales, pero para la zona norte y el centro de la ciudad aumentó los valores del indicador.

En cuanto a la variación de precios por m2 (ver **Figura 7**), los mayores crecimientos porcentuales se registraron en Villa Estrella, Providencia, Chipre, El Carmen y La Esmeralda II, barrios tradicionalmente residenciales que comienzan a mostrar signos de transformación urbana. En total, el 26% de los barrios (38 en total) presentaron un aumento en el valor del suelo por encima del promedio de la muestra, concentrándose principalmente en el norte y centro de la ciudad, áreas que reflejan una mayor presión inmobiliaria y procesos de revalorización más intensos.

Figura 7. Cambio en el Precio por m2



Fuente: elaboración de los autores con base en DANE.

4.5 La estimación del modelo Random Forest

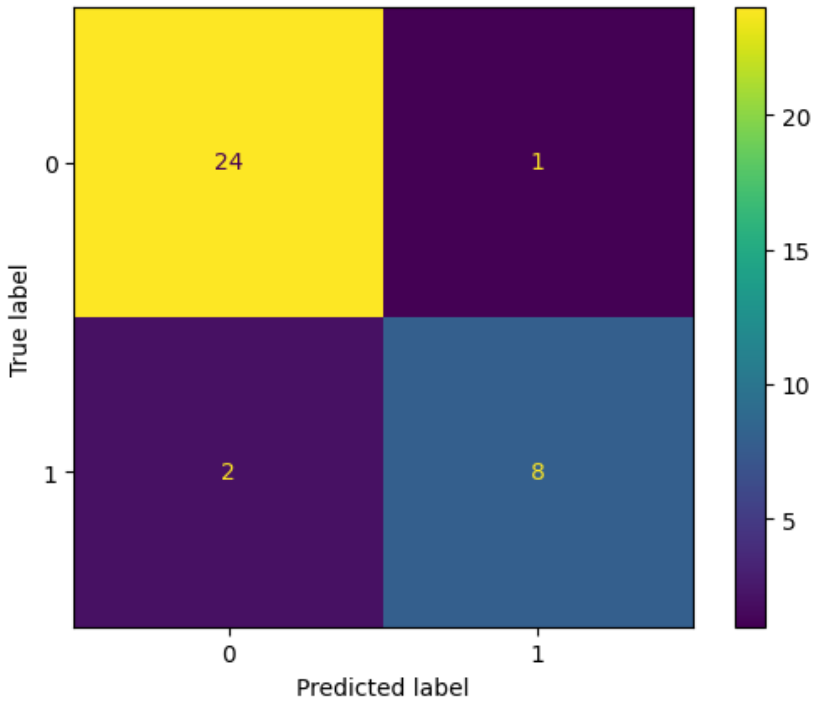
En este apartado se muestra el modelo de Random Forrest, su hiperparametrización, los indicadores de calidad del modelo, variables de importancia, y finalmente los resultados del modelo. La **Tabla 2** muestra los umbrales del clúster considerado, mientras que la **Figura 8** muestra la matriz de confusión resultante del modelo:

Tabla 2. Umbrales del clúster

Vecindarios	Número
Gentrificado (Prob>=51)	38
Gentrificable (0.10<=Prob<0.51)	27
No Gentricado (Prob<0.15)	74

Fuente: elaboración de los autores.

Figura 8. Matriz de Confusión.



Fuente: Elaboración de los autores.

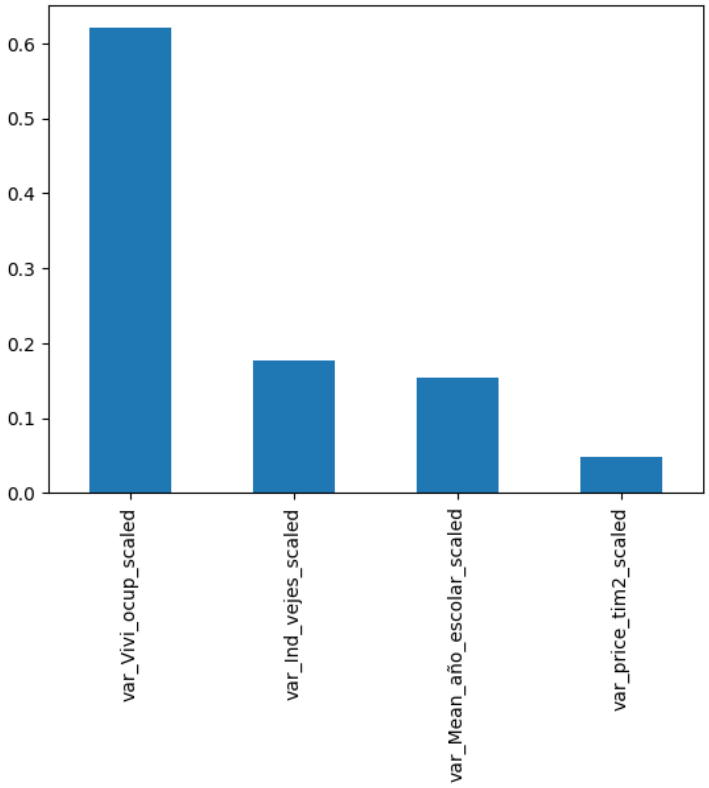
Una matriz de confusión tiene los siguientes elementos para evaluar que tan acertado es un modelo:

- **Verdadero Positivo (TP):** El modelo predice positivo y la etiqueta real es positiva.
- **Falso Negativo (FN):** El modelo predice negativo, pero la etiqueta real es positiva.
- **Falso Positivo (FP):** El modelo predice positivo, pero la etiqueta real es negativa.
- **Verdadero Negativo (TN):** El modelo predice negativo y la etiqueta real es negativa.

Se puede observar que el error del modelo es bajo, principalmente por los Falso negativos y Falsos positivos suman solo 3 observaciones en los datos de testeo.

Por otra parte, **Figura 9** presenta la importancia de las variables de modelo RF, demostrando que la variable más relevante que seleccionaron los árboles de decisión es la proporción de viviendas ocupadas escaladas. En este caso no se tiene certeza que dicha variable esta discriminada por residentes nuevos o viejos, pero presenta incidencia en la construcción de la variable de "Gentrificación" y en la predicción de los resultados.

Figura 9. Variables de importancia del RF



Fuente: Elaboración de los autores.

En la **Tabla *3*** se presentan los hiperparámetros utilizados para optimizar el modelo de Random Forest mediante *RandomizedSearchCV,* por su parte la ***Tabla 4*** muestra los parámetros de calidad del modelo escogido.

Tabla 3. Hiperparámetros del RF

Hiperparámetro	Valor
max_depth	6
min_samples_leaf	4
min_samples_split	9
n_estimators	230

Fuente: Elaboración de los autores.

Tabla 4. Calidad del Modelo

Indicador	Puntaje
Accuracy	0.9142
F1-Score	0.842
Precision	0.88
Recall	0.8

Fuente: Elaboración de los autores.

La técnica *RandomizedSearchCV* (ver ***Tabla 3***), contribuye a reducir el sobreajuste del modelo al explorar diferentes combinaciones de parámetros de forma aleatoria. La selección del mejor conjunto de hiperparámetros se realiza a través de validación cruzada con k-folds, empleando 5 particiones, con el fin de maximizar el rendimiento del modelo. Se utiliza la métrica *accuracy* como medida de desempeño y se ejecuta el proceso con *RandomizedSearchCV* de la librería scikit-learn, que identifica automáticamente los parámetros óptimos dentro del espacio de búsqueda definido. En el modelo de Random Forest, los hiperparámetros seleccionados cumplen un papel crucial en el control de la complejidad y la capacidad de generalización. El parámetro max_depth = 6 limita la profundidad máxima de los árboles, evitando que crezcan demasiado y se sobreajusten a los datos. Por su parte, min_samples_leaf = 4 establece que cada hoja debe contener al menos cuatro muestras, lo que reduce la creación de nodos terminales basados en muy pocos datos y mejora la estabilidad del modelo. El hiperparámetro min_samples_split = 9 exige un mínimo de nueve observaciones para dividir un nodo, previniendo particiones innecesarias y disminuyendo la variabilidad. Finalmente**,** n_estimators = 230 define el número de árboles que conforman el bosque, aumentando la precisión y la robustez del modelo al promediar más predicciones, aunque con un mayor costo computacional. En conjunto, estos valores permiten obtener un Random Forest equilibrado, menos propenso al sobreajuste y con mejor capacidad predictiva.

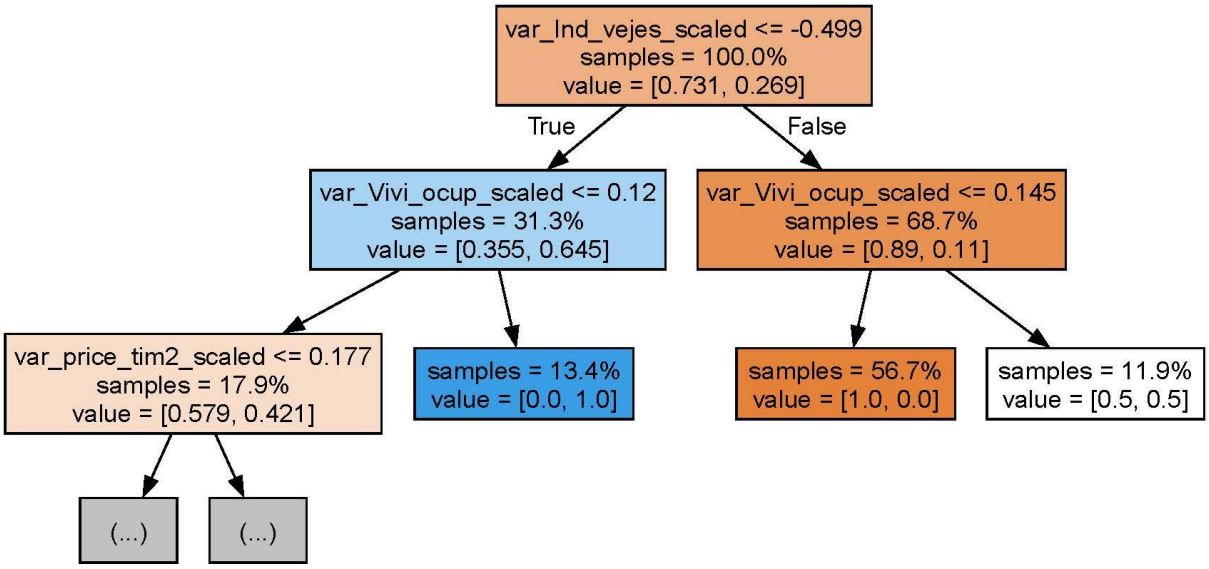
El desempeño predictivo del modelo, presentado en la **Tabla 4**, valida la robustez del enfoque metodológico adoptado. Con un accuracy de 0.9142 (91.4%), el modelo demuestra alta capacidad para clasificar correctamente los barrios en sus categorías correspondientes. La precisión de 0.88 indica que el 88% de los barrios clasificados como gentrificados efectivamente presentan características consistentes con este proceso, mientras que el recall de 0.8 revela que el modelo identifica correctamente el 80% de los barrios que realmente experimentaron gentrificación. El F1-score de 0.842 representa un equilibrio robusto entre precisión y recall, confirmando que el modelo no sacrifica la capacidad de detección por reducir falsos positivos, ni viceversa. Estos resultados son consistentes con las métricas de la matriz de confusión presentada en la **Figura 8**, donde se observó que los errores de clasificación (falsos negativos y falsos positivos) sumaron únicamente 3 observaciones en el conjunto de testeo, correspondiendo a una tasa de error del 8.6%. Este nivel de precisión predictiva es comparable, e incluso superior, a los reportados en estudios internacionales similares: el modelo desarrollado por Reades et al. (2019) para Londres alcanzó un AUC-ROC de 0.747, mientras que el estudio de gentrificación en Washington D.C. reportó un accuracy de 0.83.

4.6 Análisis de árboles de decisión individuales

Para comprender en profundidad el funcionamiento del modelo Random Forest y las reglas de decisión aprendidas, se analizaron tres árboles individuales representativos del ensamble (ver **Figura 10, 11 y 12**). Los árboles analizados presentan estructuras diversas, característica fundamental del método Random Forest que contribuye a reducir el sobreajuste y mejorar la capacidad de generalización del modelo, de esta forma, cada árbol muestra cómo el algoritmo divide recursivamente el espacio de características para clasificar los barrios en gentrificados y no gentrificados.

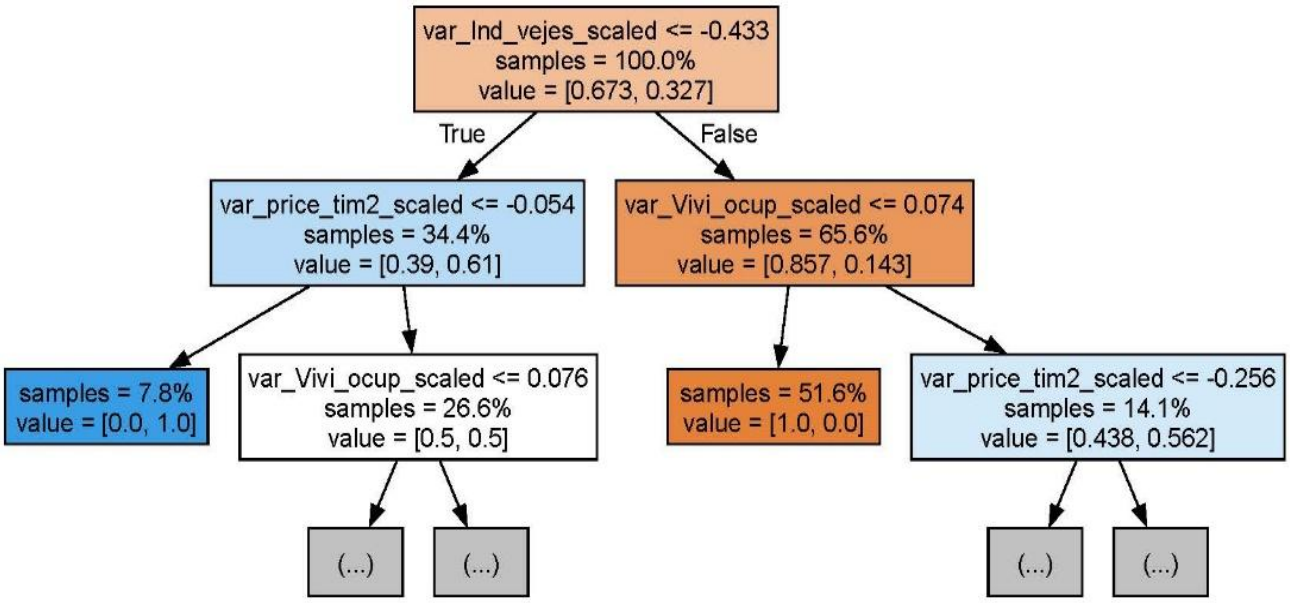
El primer árbol analizado (**Figura 10**) utiliza la variable de viviendas ocupadas como nodo raíz, identificando que valores superiores a 0.319 en la escala normalizada clasifican directamente el 16.7% de los barrios como gentrificados con un 100% de certeza. Este resultado sugiere que una alta ocupación de viviendas constituye un indicador muy robusto de gentrificación, reflejando la presión inmobiliaria característica de este fenómeno urbano. Para el 83.3% de los barrios restantes, que presentan niveles menores de ocupación, el árbol evalúa el índice de vejez como segunda variable de división, requiriendo la combinación de múltiples características para determinar la clasificación final.

Figura 10. Árbol de decisión 1 del modelo Random Forest



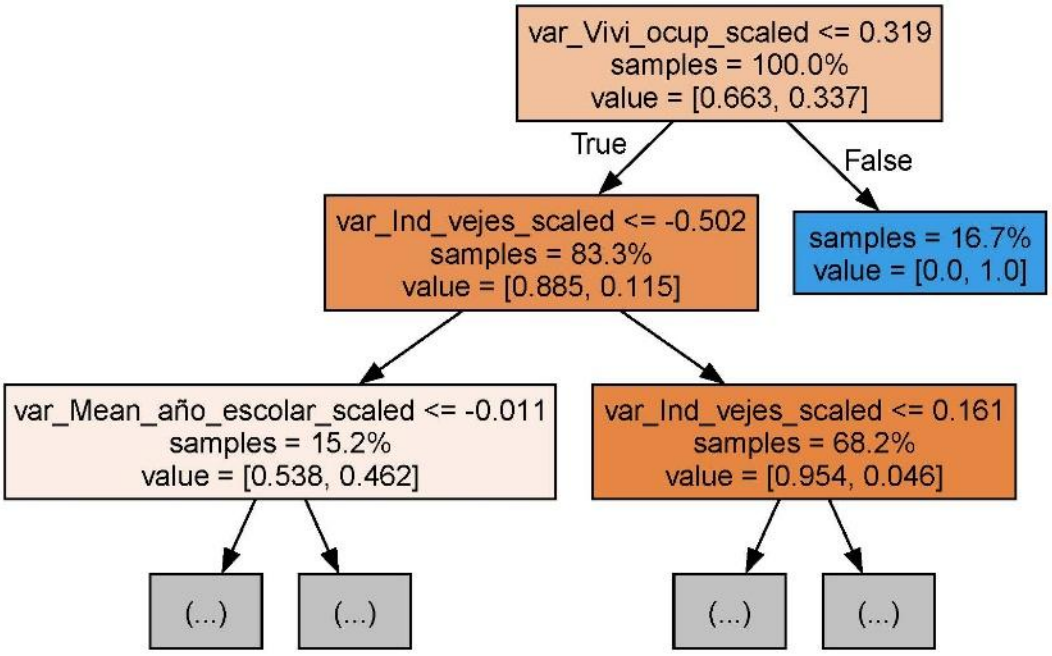
Fuente: Elaboración de los autores.

Figura 11. Árbol de decisión 2 del modelo Random Forest



Fuente: Elaboración de los autores.

Figura 12. Árbol de decisión 3 del modelo Random Forest



Fuente: Elaboración de los autores.

En contraste con el primer árbol, el segundo y tercer árbol (**Figura 11 y Figura 12**) utilizan el índice de vejez como variable raíz, aunque con umbrales diferentes de -0.433 y -0.499 respectivamente. Ambos árboles identifican que una población joven constituye el primer discriminador para predecir gentrificación, lo que refleja el rejuvenecimiento demográfico ampliamente documentado en procesos de transformación urbana. El segundo árbol clasifica el 7.8% de los barrios con población muy joven y precios del suelo estables o crecientes como gentrificados, mientras que el 51.6% de los barrios con población envejecida y baja ocupación de viviendas son clasificados como no gentrificados. Por su parte, el tercer árbol presenta un umbral más estricto para el índice de vejez (-0.499 comparado con -0.433), requiriendo una población significativamente más joven para iniciar la predicción de gentrificación.

Esta estrategia más conservadora reduce los falsos positivos, clasificando un 13.4% de los barrios como gentrificados en lugar del 7.8% del segundo árbol. La diferencia en umbrales entre estos dos árboles ilustra cómo Random Forest genera diversidad en el ensamble, permitiendo que diferentes árboles capturen matices distintos del mismo fenómeno.

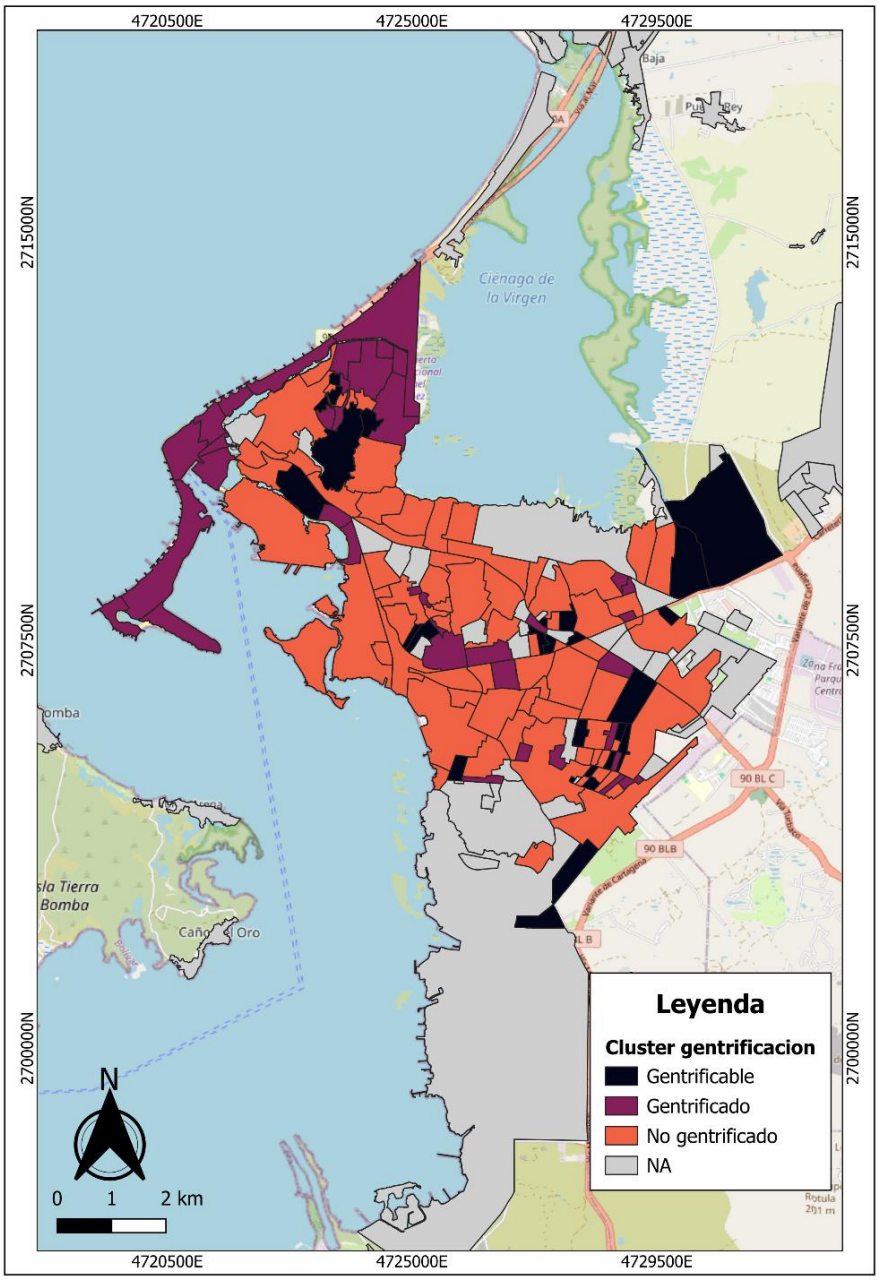
El análisis comparativo de los tres árboles revela una jerarquía clara de variables predictoras que coincide con los principios teóricos de la gentrificación. El índice de vejez aparece en los tres árboles como variable de división, confirmando su papel como el predictor más importante del modelo; valores bajos del índice, indicando población joven, están consistentemente asociados con barrios gentrificados, mientras que valores altos, reflejando población envejecida, predicen barrios no gentrificados. Este hallazgo es coherente con la teoría de gentrificación que documenta el rejuvenecimiento demográfico como un marcador fundamental del proceso, donde nuevos residentes jóvenes de clase media-alta desplazan gradualmente a la población envejecida original. La variable de viviendas ocupadas también aparece en los tres árboles, aunque con comportamientos no lineales que dependen del contexto de otras variables. En algunos casos, una alta ocupación de viviendas indica gentrificación, particularmente cuando se combina con población joven, mientras que en otros contextos puede asociarse con barrios tradicionales no gentrificados caracterizados por alta densidad residencial histórica. Esta complejidad refleja las múltiples dinámicas del mercado inmobiliario cartagenero, donde tanto la gentrificación residencial como la conversión a alojamientos turísticos pueden incrementar la ocupación de viviendas.

El precio del suelo aparece en dos de los tres árboles como variable complementaria que refina la clasificación en casos ambiguos. Su presencia es particularmente relevante cuando el índice de vejez y las viviendas ocupadas por sí solos no proporcionan señales suficientemente claras. Esto sugiere que la valorización del suelo, aunque importante, actúa más como un indicador secundario que confirma o matiza las señales demográficas e inmobiliarias primarias. Finalmente, la media de años escolares aparece solo en el primer árbol como variable de refinamiento, sugiriendo que, aunque el nivel educativo es relevante para caracterizar el perfil socioeconómico de los residentes, tiene menor poder predictivo en comparación con las variables demográficas e inmobiliarias. Esta jerarquía de importancia es consistente con estudios previos sobre gentrificación que priorizan el análisis de cambios demográficos y transformaciones del mercado de vivienda sobre indicadores educativos.

4.7 RESULTADOS DEL MODELO RANDOM FOREST

La **Figura 13** presenta los resultados consolidados del modelo, evidenciando la clasificación de los 139 barrios analizados en tres categorías principales según su probabilidad de gentrificación: barrios gentrificados, gentrificables y no gentrificados.

Figura 13. Resultados del Modelo Random Forest



Fuente: Elaboración de los autores.

Como se observa en la **Tabla 2**, el modelo clasificó 38 barrios (26% del total) como gentrificados, es decir, vecindarios con una probabilidad superior al 51% de haber experimentado procesos consolidados de transformación urbana asociados a gentrificación. Por otra parte, 27 barrios (18.5% del total) fueron clasificados como gentrificables, presentando probabilidades de gentrificación entre 0.10 y 0.51. Las **Tabla 5 y 6** muestran estos resultados a detalle.

Tabla 5. Barrios clasificados como gentrificables por el modelo desarrollado.

Barrio	Cluster_gentrificacion
ALAMEDA LA VICTORIA	Gentrificable
ALTOS DE SAN ISIDRO	Gentrificable
BELLAVISTA	Gentrificable
CAMILO TORRES	Gentrificable
CERRO DE LA POPA	Gentrificable
EL POZON	Gentrificable
JAIME PARDO LEAL	Gentrificable
JORGE ELIECER GAITAN	Gentrificable
LA CASTELLANA	Gentrificable
LA ESMERALDA I	Gentrificable
LA FLORIDA	Gentrificable
LAS DELICIAS	Gentrificable
LOS CERROS	Gentrificable
LOS JARDINES	Gentrificable
NAZARENO	Gentrificable
PABLO VI - I	Gentrificable
PEDRO SALAZAR	Gentrificable
PETARE	Gentrificable
PIE DE LA POPA	Gentrificable
POLICARPA	Gentrificable
RUBI	Gentrificable
SAN ANTONIO	Gentrificable
SAN BERNARDO	Gentrificable
TACARIGUA	Gentrificable
VILLA ESTRELLA	Gentrificable
VILLA HERMOSA	Gentrificable
VILLA ROSITA	Gentrificable

Fuente: Elaboración de los autores.

La **Tabla 5** detalla estos vecindarios clasificados como gentrificables, entre los que destacan El Pozón, Villa Estrella, Alameda la Victoria, y varios sectores de Olaya Herrera (Pablo VI-I, Pedro Salazar, Jaime Pardo Leal, Jorge Eliécer Gaitán). Esta clasificación identifica barrios que, si bien no han experimentado transformaciones consolidadas, presentan condiciones socioeconómicas y dinámicas del mercado inmobiliario que los hacen susceptibles de gentrificación futura.

La inclusión de Villa Estrella en esta categoría es coherente con los hallazgos presentados en la **Figura 7**, donde este barrio registró uno de los mayores crecimientos porcentuales en precio del suelo por metro cuadrado durante el periodo 2005-2018.

Tabla 6. Barrios clasificados como gentrificado por el modelo desarrollado.

Barrio	cluster_gentrificacion
ALMIRANTE COLON	Gentrificado
BARRIO CHINO	Gentrificado
BOCAGRANDE	Gentrificado
BOSQUECITO	Gentrificado
CANAPOTE	Gentrificado
CASTILLOGRANDE	Gentrificado
CENTRO	Gentrificado
CRESPO	Gentrificado
DANIEL LEMAITRE	Gentrificado
EL CABRERO	Gentrificado
EL COUNTRY	Gentrificado
EL GALLO	Gentrificado
EL LAGUITO	Gentrificado
EL LIBERTADOR	Gentrificado
EL REPOSO	Gentrificado
GETSEMANI	Gentrificado
JOSE ANTONIO GALAN	Gentrificado
JUNIN	Gentrificado
LA MATUNA	Gentrificado
LA SIERRITA	Gentrificado
LOMA FRESCA	Gentrificado
LOS ANGELES	Gentrificado
LUIS CARLOS GALAN	Gentrificado
MARBELLA	Gentrificado
MARTINEZ MARTELO	Gentrificado
NUEVA DELHI	Gentrificado
NUEVA JERUSALEN	Gentrificado
NUEVO BOSQUE	Gentrificado
NUEVO PORVENIR	Gentrificado
PABLO VI - II	Gentrificado
PARAISO II	Gentrificado
SAN DIEGO	Gentrificado
SAN FRANCISCO	Gentrificado
SAN JOSE OBRERO	Gentrificado
SANTA MARIA	Gentrificado
SANTA MONICA	Gentrificado
SIETE DE AGOSTO	Gentrificado
VILLA RUBIA	Gentrificado

Fuente: Elaboración de los autores.

Estos barrios clasificados como gentrificados, detallados en la **Tabla 6**, incluyen sectores emblemáticos del centro histórico como Getsemaní, Centro y San Diego, así como zonas de la franja norte de la ciudad como Bocagrande, Castillogrande, El Laguito, Marbella y El Cabrero. La presencia de barrios como Chambacú en esta categoría resulta particularmente significativa, dado el dramático incremento del 24% en años promedio de escolaridad y la ocupación del 100% de viviendas registrada en 2018, indicadores que reflejan una transformación socioeconómica profunda del vecindario.

Finalmente, 74 barrios (50.7% del total) fueron clasificados como no gentrificados, con probabilidades inferiores al 15%, indicando vecindarios que no presentan señales significativas de transformación urbana asociada a gentrificación en el periodo analizado. La distribución espacial de los resultados revela patrones geográficos consistentes con los hallazgos descriptivos presentados en las **Figura 4, 5, 6 y 7**. Los barrios gentrificados se concentran en dos núcleos principales: el centro histórico y su área de influencia inmediata, y la zona norte costera de la ciudad. Este patrón espacial refleja las dinámicas diferenciadas de gentrificación documentadas en el marco teórico: turistificación y conversión de vivienda permanente en alojamientos de corta estancia en el centro histórico (fenómeno paradigmáticamente observado en Getsemaní), y procesos de super-gentrificación impulsados por inversión inmobiliaria de alto standing en la zona norte. Los barrios gentrificables, por su parte, muestran una distribución predominantemente periférica, con concentración en sectores tradicionalmente caracterizados por menores niveles socioeconómicos. La presencia de barrios como El Pozón en esta categoría sugiere que las presiones de mercado inmobiliario y las expectativas de revalorización están comenzando a alcanzar áreas históricamente relegadas del desarrollo urbano formal, potencialmente asociadas a proyectos de infraestructura vial, expansión urbana o especulación del suelo.

Cabe destacar que la presencia de algunos barrios tanto en la franja de gentrificados como en la de gentrificables no implica necesariamente una transición urbana asociada a la turistificación. Esto se debe a que las variables utilizadas reflejan mejoras en la calidad de vida de los residentes y avances en infraestructura social, como la educación, por lo que pueden estar respondiendo a cambios sociodemográficos sin que ello conlleve, por ejemplo, un aumento en la renta corta tipo Airbnb. No obstante, es importante resaltar aquellos barrios con potencial de crecimiento económico que podrían estar experimentando efectos de desplazamiento o verse absorbidos por la nueva ola de gentrificación que atraviesa Cartagena. Este análisis y la estimación del modelo ofrecen pistas e hipótesis que permiten identificar los vecindarios que están experimentando transformaciones en su dinámica urbana.

CONCLUSIONES

Este estudio ha propuesto un modelo de Random Forest para la identificación de barrios gentrificados y susceptibles a la gentrificación en la ciudad de Cartagena, Colombia, utilizando datos censales correspondientes a los años 2005 y 2018. La investigación representa una contribución metodológica significativa al campo de los estudios urbanos en el contexto latinoamericano, al aplicar técnicas de aprendizaje automático supervisado para el análisis de procesos de transformación urbana que tradicionalmente han sido abordados mediante enfoques predominantemente cualitativos. El modelo Random Forest desarrollado alcanzó un desempeño predictivo sobresaliente, con un accuracy del 91.4%, una precisión de 0.88, un recall de 0.8 y un F1-score de 0.842, métricas que demuestran la robustez y confiabilidad del enfoque metodológico para la identificación de procesos de gentrificación en Cartagena. De los 139 barrios analizados, el modelo clasificó 38 (26%) como gentrificados y 27 (18.5%) como gentrificables, evidenciando que aproximadamente el 44.5% de los vecindarios de la ciudad han experimentado o están en riesgo de experimentar transformaciones urbanas asociadas a procesos de revalorización inmobiliaria y cambio en la composición socioeconómica de sus residentes. La matriz de confusión revela que el modelo cometió solo 3 errores de clasificación en el conjunto de prueba, correspondiendo a una tasa de error del 8.6%, lo que confirma su capacidad para generalizar correctamente a datos no vistos durante el entrenamiento. Este nivel de precisión es comparable, e incluso superior, al reportado en estudios similares desarrollados en Washington D.C. (accuracy de 0.83), Londres (AUC-ROC de 0.747) y Sídney, posicionando al modelo como una herramienta confiable para la detección temprana de gentrificación en contextos urbanos latinoamericanos.

Los resultados del análisis descriptivo de los patrones urbanos observados entre 2005 y 2018 evidencian transformaciones significativas en la estructura socioeconómica y habitacional de Cartagena. Los barrios del centro histórico y la zona norte de la ciudad, particularmente Chambacú, El Cabrero, El Laguito, Bocagrande y Castillo Grande, presentaron los incrementos más pronunciados en indicadores asociados a procesos de gentrificación: aumento del nivel educativo promedio (con Chambacú registrando un incremento del 24%), elevación sustancial en el precio por metro cuadrado, y modificaciones en la proporción de viviendas ocupadas que sugieren procesos de turistificación y sustitución de vivienda permanente por usos temporales. Estos patrones son consistentes con la literatura sobre gentrificación transnacional en ciudades turísticas latinoamericanas, donde la competencia por el espacio urbano entre residentes originales y nuevos inversores inmobiliarios genera presiones de desplazamiento en poblaciones de bajos ingresos.

El análisis de importancia de variables mediante el método de permutación revela que la proporción de viviendas ocupadas escalada constituye el predictor más influyente en el modelo Random Forest, seguida por el índice de envejecimiento escalado, el precio del suelo por metro cuadrado en el periodo 2018 y, en menor medida, la media de años de escolaridad. Este ordenamiento jerárquico de variables es coherente con los marcos teóricos de gentrificación que identifican tres dimensiones fundamentales del fenómeno: transformaciones en el mercado de vivienda (reflejadas en ocupación y precio), cambios demográficos (capturados por el rejuvenecimiento poblacional), y mejoras en el capital

humano de los residentes (evidenciadas en el nivel educativo). La preeminencia de la proporción de viviendas ocupadas como variable más importante sugiere que en el contexto cartagenero, las dinámicas de turistificación y conversión de vivienda permanente en alojamientos de corta estancia (fenómeno documentado paradigmáticamente en Getsemaní, donde la población residente se redujo de 10.500 en 2005 a 2.300 en 2018, y apenas 448 en 2025) constituyen un marcador particularmente sensible de gentrificación. Por su parte, el índice de envejecimiento emerge como predictor crítico debido a que el rejuvenecimiento demográfico representa un síntoma temprano y consistente de transformación barrial, mientras que el precio del suelo actúa como variable complementaria que confirma procesos ya iniciados más que como indicador anticipatorio.

La adaptación metodológica realizada para el contexto de Cartagena, que incorpora el índice de envejecimiento en sustitución de variables sobre ingresos de hogares (no disponibles en el censo colombiano), demuestra la viabilidad de aplicar modelos predictivos de gentrificación en contextos donde la disponibilidad de datos es limitada. La construcción de la variable dependiente mediante cinco criterios socioeconómicos (variación de viviendas, viviendas ocupadas, años de escolaridad, precio del suelo, e índice de envejecimiento) permite capturar de forma multidimensional los procesos de cambio urbano, superando las limitaciones de aproximaciones unidimensionales que se basan exclusivamente en indicadores de ingreso o precio de vivienda. La capacidad predictiva del modelo Random Forest se extiende más allá de la clasificación retrospectiva del periodo 2005-2018, ofreciendo potencial para la extrapolación temporal y espacial de procesos de gentrificación en Cartagena. La arquitectura del modelo, con 230 árboles de decisión y hiperparámetros optimizados mediante RandomizedSearchCV (max_depth=6, min_samples_split=9, min_samples_leaf=4), configura un equilibrio óptimo entre capacidad de aprendizaje y prevención de sobreajuste, permitiendo que el modelo capture patrones generalizables. En términos de extrapolación temporal, el modelo puede actualizarse con datos censales del próximo ciclo (proyectado para 2025-2026) para generar predicciones sobre el estado de gentrificación hacia 2030, permitiendo identificar barrios que transitarán de la categoría 'gentrificable' a 'gentrificado', así como detectar nuevos vecindarios en riesgo. Desde una perspectiva espacial, el modelo entrenado en Cartagena podría adaptarse mediante transfer learning a otras ciudades costeras colombianas con dinámicas turísticas similares (Santa Marta, Barranquilla) o ciudades intermedias latinoamericanas que experimentan presiones inmobiliarias análogas. No obstante, el modelo asume que las relaciones entre variables predictoras y gentrificación observadas en 2005-2018 se mantendrán estables, supuesto que puede violarse ante cambios drásticos en política urbana o shocks económicos, recomendándose implementar un sistema de monitoreo continuo que recalibre el modelo periódicamente.

El uso de algoritmos de aprendizaje automático, particularmente Random Forest, presenta ventajas metodológicas sustanciales para el estudio de la gentrificación urbana. La capacidad del modelo para manejar relaciones no lineales complejas entre variables, identificar interacciones entre factores socioeconómicos sin especificación previa, y proporcionar medidas de importancia de variables, permite una comprensión más profunda de los mecanismos subyacentes a los procesos de transformación urbana. Además, la posibilidad de realizar predicciones sobre barrios susceptibles a la gentrificación (nowcasting) representa una herramienta valiosa para la planificación urbana preventiva,

permitiendo a los responsables de política pública anticipar procesos de desplazamiento antes de que se consoliden.

No obstante, es importante reconocer las limitaciones del presente estudio. En primer lugar, la disponibilidad limitada de datos censales en Colombia (con actualizaciones al menos cada 10 años, aunque con periodos de tiempo irregulares) restringe la capacidad de capturar dinámicas de cambio urbano de corto y mediano plazo, particularmente relevantes en el contexto de la rápida turistificación experimentada por Cartagena en la última década. En segundo lugar, la ausencia de información sobre ingresos de los hogares, composición étnico-racial de los barrios, y patrones de tenencia de vivienda (propiedad vs. arrendamiento) limita la capacidad del modelo para capturar dimensiones importantes del proceso de gentrificación documentadas en la literatura internacional. En tercer lugar, el estudio se centra exclusivamente en variables socioeconómicas y habitacionales, sin incorporar explícitamente dimensiones espaciales como proximidad a amenidades urbanas, accesibilidad al transporte público, o transformaciones en el uso del suelo, que la literatura ha identificado como factores relevantes en procesos de gentrificación.

Si bien el modelo Random Forest demuestra un desempeño robusto, es importante reconocer limitaciones metodológicas específicas. En primer lugar, el tamaño relativamente pequeño de la muestra (139 barrios, de los cuales 35 se utilizaron para testeo) introduce incertidumbre en las estimaciones de desempeño, especialmente para la clase minoritaria (barrios gentrificados). Técnicas de aumento de datos mediante simulación de vecindarios sintéticos o incorporación de datos de ciudades comparables podrían robustecer las estimaciones. En segundo lugar, aunque la validación cruzada con 5 folds mitiga riesgos de sobreajuste, el modelo no captura heterogeneidad en tipos de gentrificación: no distingue entre gentrificación residencial tradicional, super-gentrificación en barrios ya acomodados, turistificación, y 'new-build gentrification' asociada a grandes proyectos inmobiliarios. En tercer lugar, la ausencia de variables espaciales explícitas (proximidad a amenidades, accesibilidad a transporte, distancia al CBD) que la literatura identifica como predictores importantes podría limitar la capacidad explicativa del modelo. La incorporación de estas variables mediante análisis de redes o modelos espacialmente explícitos (spatial lag, spatial error) podría incrementar tanto la precisión predictiva como la interpretabilidad. Finalmente, el modelo no incorpora información temporal de alta frecuencia sobre transacciones inmobiliarias, licencias turísticas o datos de plataformas como Airbnb, que podrían detectar señales tempranas de transformación antes de que sean capturadas en censos decenales.

RECOMENDACIONES

Los hallazgos de esta investigación tienen implicaciones significativas para la formulación de políticas urbanas en Cartagena y otras ciudades intermedias latinoamericanas que experimentan procesos similares de turistificación y revalorización inmobiliaria. Se recomienda a los responsables de política pública considerar las siguientes acciones:

- Implementar sistemas de monitoreo continuo de los indicadores socioeconómicos y habitacionales identificados en este estudio como predictores de gentrificación, con el fin de desarrollar capacidades de detección temprana de procesos de transformación urbana que puedan derivar en desplazamiento de poblaciones vulnerables.
- Diseñar e implementar instrumentos de regulación del mercado inmobiliario y turístico que permitan equilibrar los objetivos de desarrollo económico con la protección del derecho a la vivienda de los residentes originales, incluyendo mecanismos de control de alquileres, restricciones al uso de vivienda para fines turísticos de corta estancia, e incentivos para la producción y preservación de vivienda asequible en barrios en proceso de revalorización.
- Desarrollar estrategias de renovación urbana que no impliquen desplazamiento social, incorporando a las comunidades residentes en los procesos de toma de decisiones sobre intervenciones urbanas y garantizando su permanencia mediante políticas de vivienda social, mejoramiento barrial participativo, y fortalecimiento del tejido organizativo local.
- Fortalecer los sistemas de información urbana y estadística oficial, incorporando variables adicionales en los censos de población y vivienda que permitan caracterizar de forma más completa los procesos de cambio urbano, incluyendo información sobre ingresos de los hogares, composición étnico-racial, patrones de tenencia, y movilidad residencial.
- Promover investigaciones interdisciplinarias que complementen los enfoques cuantitativos con metodologías cualitativas, permitiendo una comprensión más profunda de las experiencias vividas por las comunidades afectadas por procesos de gentrificación y las estrategias de resistencia desarrolladas por los residentes originales.

Los resultados del modelo permiten formular recomendaciones diferenciadas según el nivel de riesgo de gentrificación identificado. Para los 27 barrios clasificados como gentrificables (El Pozón, Villa Estrella, Alameda la Victoria, sectores de Olaya Herrera, entre otros), se recomienda implementar estrategias de intervención preventiva que incluyan: (1) establecimiento de zonas de preservación residencial con restricciones al uso turístico comercial; (2) creación de fondos de adquisición de suelo para bancos de tierra pública que permitan al Estado anticiparse a la especulación; (3) implementación de mecanismos de 'community land trusts' que transfieran propiedad colectiva a organizaciones barriales; y (4) fortalecimiento del tejido organizativo local mediante apoyo a juntas de acción comunal y asociaciones de vecinos. Para los 38 barrios ya gentrificados (Getsemaní, Centro, San Diego, Bocagrande, Castillogrande, El Laguito, Chambacú, entre otros), donde procesos de desplazamiento están consolidados, se recomiendan políticas de mitigación que incluyan: (1) regulación estricta de plataformas de alquiler de corta estancia mediante cuotas

máximas de licencias turísticas por barrio; (2) implementación de control de alquileres (rent control) especialmente en el centro histórico; (3) creación de programas de 'derecho al retorno' para familias desplazadas; (4) desarrollo de políticas de vivienda inclusiva (inclusionary zoning) que exijan destinar 20-30% de nuevas unidades a vivienda asequible; y (5) implementación de impuestos redistributivos sobre valorización del suelo. Transversalmente, se recomienda establecer un observatorio urbano permanente que utilice el modelo Random Forest como herramienta de nowcasting para monitorear trimestralmente indicadores clave y actualizar clasificaciones de riesgo.

Futuras investigaciones deberían explorar la incorporación de fuentes de datos no tradicionales, como información de plataformas de alquiler de corta estancia (Airbnb), transacciones inmobiliarias de alta frecuencia, imágenes satelitales para detectar cambios en el entorno construido, y datos de redes sociales que permitan capturar percepciones y narrativas sobre los procesos de transformación urbana. Asimismo, resulta pertinente evaluar la aplicabilidad del modelo propuesto en otras ciudades colombianas y latinoamericanas que experimentan procesos similares de turistificación y presión inmobiliaria, con el fin de validar su capacidad de generalización y adaptabilidad a distintos contextos urbanos.

REFERENCIAS BIBLIOGRÁFICAS

1. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
2. Casali, Y., Aydin, N. Y., & Comes, T. (2022). Machine learning for spatial analyses in urban areas: a scoping review. In *Sustainable Cities and Society* (Vol. 85). Elsevier Ltd. https://doi.org/10.1016/j.scs.2022.104050
3. Castro, D., Alejandro, C., & Fernanda Alejandro, M. (2020). *El proceso de Gentrificación, intervención urbana arquitectónica en la ciudad de Salinas-Ecuador*. https://www.redalyc.org/
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785
5. Delgado, J. P., Bryam, V., Guachichullca Zhiña, A., Adrián, B., & Alvear, O. (2025). *Identificación de la gentrificación urbana en base de un modelo predictivo en la ciudad de Valparaíso*. https://dspace.ucuenca.edu.ec/
6. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Fernández-Delgado, A. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? In *Journal of Machine Learning Research* (Vol. 15). http://www.mathworks.es/products/neural-network.
7. Jun, M. J. (2021). A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area. *International Journal of Geographical Information Science*, *35*(11), 2149–2167. https://doi.org/10.1080/13658816.2021.1887490
8. Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, *6*(1). https://doi.org/10.1186/1758-2946-6-10
9. Lerena-Rongvaux, N. (2023). ¿Renovación sin gentrificación? Hacia un abordaje crítico de procesos urbanos excluyentes en América Latina. Casos en Buenos Aires. *Eure*, *49*(146). https://doi.org/10.7764/eure.49.146.08

10. Loukaitou-Sideris, A., Gonzalez, S., & Ong, P. (2019). Triangulating Neighborhood Knowledge to Understand Neighborhood Change: Methods to Study Gentrification. *Journal of Planning Education and Research*, *39*(2), 227–242. https://doi.org/10.1177/0739456X17730890
11. Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. http://arxiv.org/abs/1705.07874
12. Mansour, Y., & Schain, M. (2001). Learning with Maximum-Entropy Distributions. *Machine Learning*, *45*(2), 123–145. https://doi.org/10.1023/A:1010950718922
13. Maya, M., Simi, H., & Pearsall, H. (2024). Machine learning to model gentrification: A synthesis of emerging forms. In *Computers, Environment and Urban Systems* (Vol. 111). Elsevier Ltd. https://doi.org/10.1016/j.compenvurbsys.2024.102119
14. Perren, J., & Cabezas, S. (2018). ¿Gentrificación en el “fin del mundo”? Crecimiento en altura y elitización en una ciudad intermedia de la Patagonia (Neuquén, 2001-2010). *QUID 16: Revista Del Área De Estudios Urbanos Del Instituto De Investigaciones GINO GERMANI De La Facultad De CIENCIAS SOCIALES (UBA)*.
15. Peter Bühlmann, B. (2002). ANALYZING BAGGING. In *The Annals of Statistics* (Vol. 30, Issue 4).
16. Probst, P., Wright, M., & Boulesteix, A.-L. (2019). *Hyperparameters and Tuning Strategies for Random Forest*. https://doi.org/10.1002/widm.1301
17. Quinlan, J. R. (1993). *C4.5: PROGRAMS FOR MACHINE LEARNING* (P. Langley, Ed.). Morgan Kaufmann Publishers, Inc.
18. Reades, J., De Souza, J., & Hubbard, P. (2019). Understanding urban gentrification through machine learning. *Urban Studies*, *56*(5), 922–942. https://doi.org/10.1177/0042098018789054
19. Salinas Arreortua, L. A. (2013). Gentrificación en la ciudad latinoamericana: el caso de Buenos Aires y Ciudad de México. *GeoGraphos. Revista Digital Para Estudiantes de Geografía y Ciencias Sociales*, *4*. https://doi.org/10.14198/geogra2013.4.44
20. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms* (First edition). Cambridge University Press. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

21. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-307
22. Thackway, W., Ng, M., Lee, C. L., & Pettit, C. (2023). Building a predictive machine learning model of gentrification in Sydney. *Cities*, *134*. https://doi.org/10.1016/j.cities.2023.104192
23. The MathWorks, Inc. (2016). *Introducing Machine Learning*.
24. Villanueva, C. L., & Vallbona, M. C. (2021). Gentrificación y turistificación: dinámicas y estrategias en Barcelona. *Encrucijadas. Revista Crítica de Ciencias Sociales, 21(1), A2102.*, *21*(1), 2102.
25. Yee, J., & Dennett, A. (2022). Stratifying and predicting patterns of neighbourhood change and gentrification: An urban analytics approach. *Transactions of the Institute of British Geographers*, *47*(3), 770–790. https://doi.org/10.1111/tran.12522
26. Yoo, J., & Census Bureau, U. S. (2023). *Identifying Gentrification using Machine Learning **. https://www.census.gov/programs-surveys/ahs/tech-documentation/def-errors-changes.html
27. Rodríguez Mejía, J. L. Ciudades y globalización: capitalismo de plataformas y gentrificación en Nueva York, Londres y Ciudad de México (2008-2023) (Master's thesis, Quito, Ecuador: Flacso Ecuador).
28. Uribe, D. B. (2024). El realismo trágico de la gentrificación en Cartagena de Indias: el caso Getsemaní. Revista Aläula, 7, 83-85.
29. Chapple, K., Waddell, P., Chatman, D., Zuk, M., Loukaitou-Sideris, A., Ong, P., ... & Gonzalez, S. R. (2017). Developing a new methodology for analyzing potential displacement.
30. Instituto de Políticas Públicas, Regional y de Gobierno. (2025). Diásporas y resistencias: Resultados del censo de población y de vivienda en Getsemaní, 2025. Universidad de Cartagena.