Motor de Inteligencia de Mercado para el Sector de Techos en Miami-Dade

De Datos Públicos Dispersos a Estrategias de Captación Accionables

Luis Aurelio Mendoza Michel

Diciembre 2025

Resumen

Este trabajo propone un Índice de Demanda de Reemplazo de Techo (RDI) para priorizar parcelas residenciales en el condado de Miami-Dade. Se integraron diversas fuentes públicas (catastro, permisos de construcción, violaciones de código y directorios de contratistas) en un master dataset unificado de 642,539 parcelas.A partir de estos datos, estimamos el año de último techo (LastRoofYear) combinando jerárquicamente la fecha del permiso más reciente y el año de construcción, ponderando cada fuente mediante una Puntuación de Confianza continua en el intervalo $[0, 1]$ para cuantificar la incertidumbre. Con esta base, construimos el RDI integrando la edad relativa del techo, su proximidad al fin de vida útil normativa, el historial de violaciones y la saturación de contratistas por código postal.Finalmente, se utilizó un modelo de regresión logística exploratorio para validar la capacidad predictiva de estas variables sobre la actividad de permisos recientes. Los resultados permiten segmentar geográficamente el mercado, identificando zonas de alta demanda latente y baja oferta competitiva para optimizar estrategias de captación comercial.

1. Introducción

En las últimas décadas, la gestión de techos residenciales en el sur de Florida ha evolucionado de un problema de mantenimiento rutinario a un desafío sistémico de vulnerabilidad urbana. Estudios recientes señalan que la convergencia de un parque habitacional envejecido —especialmente aquellas estructuras construidas antes del código post-Andrew de 1992— y la exposición recurrente a huracanes, coloca a condados como Miami-Dade en una situación de alto riesgo estructural [1].

A este escenario físico se suma una crisis financiera: el mercado de seguros en Florida se ha vuelto excepcionalmente restrictivo. Según reportes de la industria, las primas de seguros en la región han aumentado desproporcionadamente en comparación con el promedio nacional, impulsadas en gran parte por litigios relacionados con reclamos de techos y fraude [2]. En este contexto, la decisión de dónde focalizar inversiones en reemplazo de techos deja de ser un problema técnico para convertirse en un imperativo de inteligencia de mercado.

Actualmente, los gobiernos locales y actores privados disponen de múltiples fuentes de datos (catastros, permisos, violaciones), pero estas suelen estar fragmentadas.La carencia

de una métrica unificada que estime el "último año de techo" (*Last-Roof-Year*) impide una priorización eficiente de recursos.

Este trabajo aborda dicho vacío mediante la construcción de un **Índice de Demanda de Reemplazo de Techo (RDI)**. La hipótesis central es que la integración de señales débiles —como la antigüedad de la estructura y vacíos en el historial de permisos— permite inferir la vulnerabilidad de una propiedad con mayor precisión que el uso de variables aisladas.

2. Arquitectura de Datos y Metodología de Integración

2.1. Auditoría y Caracterización de Fuentes Heterogéneas

El punto de partida del proyecto fue la auditoría de un inventario de fuentes públicas del condado de Miami-Dade. Se evaluaron conjuntos de datos relacionados con el parque habitacional y la actividad de construcción bajo criterios de **completitud, consistencia temporal y relevancia predictiva**.

Tras el filtrado inicial, se seleccionaron cinco fuentes primarias para la construcción del *master dataset*. A continuación se detalla su caracterización técnica y rol estratégico en el modelo:

Cuadro 1: Matriz de Auditoría Técnica de Fuentes de Datos

Fuente	URL / Origen	Tipo	Extracción	Campos Claves	Act.	Dificultad
Catastro Parcels	**GIS MDC Open Data** (Capa 'Parcel')	GIS + CSV	Descarga directa del portal y carga en Python con GeoPandas.	PID/FOLIO, address, zip_code, year_built, building_age.	Anual	**Alta:** Alto volumen, limpieza de FOLIO, estandarización.
Building Permits	**Miami-Dade Open Data** ('Building Permits Issued')	CSV	Descarga directa y filtrado en Python (residenciales y techo).	PROCNUM, FOLIO, PermitType, ISSUDATE, ESTVALUE.	Semanal	**Media-Alta:** Filtrado tipología, manejo fechas, deducción techo.
Building Violations	**Miami-Dade Open Data** ('Building Violations')	CSV	Descarga CSV y agregación en Pandas por FOLIO.	FOLIO, X, Y, VIOL_NAME, CASE_TYPE, OPEN_DATE.	Continua	**Media:** Limpieza fechas y cálculo años desde última sanción.
Roofing Contrac.	**Florida DBPR** (Portal de Licencias)	CSV	Descarga CSV, depuración y agregación por ZIP.	contractor_name, address, zip_code, n_contractors.	Periódica	**Baja-Media:** Limpieza direcciones y conteo por ZIP.
ACS 2021	**Census.gov Release** (Data 2021)	JSON / CSV	**Conexión vía API:** Peticiones HTTP (GET) con API Key y conversión a DataFrame.	GEOID_BG, total_pop, median_income, poverty_rate.	Estát. (2021)	**Media:** Manejo de requests, paginación y estructura JSON.

Estratificación del Uso de Datos

Basado en la auditoría, se definieron tres niveles de utilización para las fuentes:

- **Nivel 1 - Integración Core:** Catastro, Permisos, Violaciones y Contratistas. Estas fuentes se cruzaron a nivel de parcela para alimentar directamente el algoritmo del RDI.
- **Nivel 2 - Exclusión por Calidad:** Fuentes como reportes al 311 y capas FEMA fueron descartadas por inconsistencias críticas en la normalización de direcciones
- **Nivel 3 - Contexto Descriptivo (ACS 2021):** Aunque se auditó y extrajo exitosamente vía API, esta fuente **no se integró al *master dataset*** parcelario. Se utilizó

exclusivamente para caracterización territorial agregada, ya que la vinculación espacial fina (Parcela → *Census Block Group*) quedó fuera del alcance de esta iteración..

.

2.2. Arquitectura del Pipeline de Integración (ETL)

Para consolidar las fuentes heterogéneas descritas en la auditoría, se diseñó e implementó una arquitectura de procesamiento ETL (*Extract, Transform, Load*) basada en Python. El flujo de trabajo, ilustrado en la **Figura 1**, transforma datos crudos (CSV transaccionales y polígonos GIS) en una estructura unificada y geoespacialmente coherente.

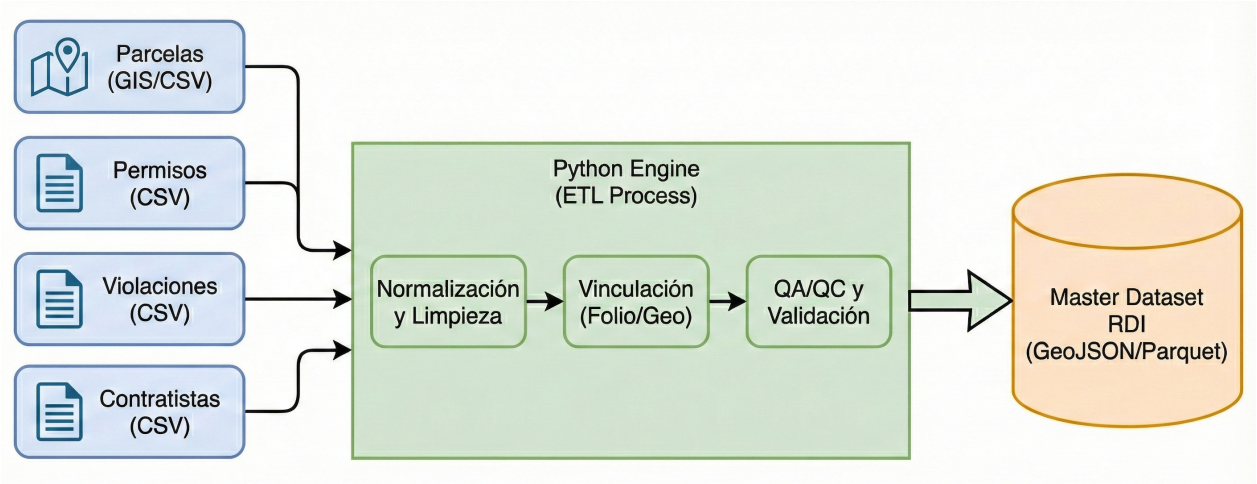

Figura 1: **Arquitectura del Sistema de Integración:** Flujo desde la ingestión de fuentes primarias (CSV/GIS) hasta la generación del Master Dataset geoespacial.

El proceso se articula en tres etapas lógicas que garantizan la integridad de los datos:

1. **Ingestión y Normalización:** Se ingieren cuatro flujos de datos principales (Parcelas, Permisos, Violaciones y Contratistas). En esta etapa, el motor de Python estandariza tipologías de datos y homologa las claves primarias (e.g., limpieza de `FOLIO` y `ZIP Code`), eliminando la fricción sintáctica entre formatos.

2. **Vinculación Relacional y Espacial:** Se utiliza la capa de **Parcelas (GIS)** como entidad ancla (*Anchor Entity*). Sobre esta base geométrica se ejecutan operaciones de *join* determinístico para atributos físicos y regulatorios, y agregaciones espaciales para la oferta de contratistas.

3. **Consolidación y QA:** El resultado es el *Master Dataset RDI*, un archivo geoespacial optimizado donde cada fila representa una propiedad única enriquecida con su historial de riesgo y mantenimiento. Este dataset final valida la completitud espacial (75 % de cobertura efectiva) antes de alimentar el modelo de decisión.

3. Diseño del Motor de Inteligencia y Validación Analítica

Tras la consolidación del *dataset* espacial, se diseñó un motor de inferencia en dos etapas secuenciales. La primera fase reconstruye el historial de mantenimiento de cada propiedad utilizando una heurística determinista con gestión de incertidumbre. La segunda fase sintetiza esta información en un índice de mercado (RDI) para la priorización de recursos.

3.1. Algoritmo de Inferencia Temporal (T_{roof})

Debido a la dispersión de los registros, se diseñó una heurística determinista para inferir la variable latente T_{roof} (Año del Último Techo). El algoritmo selecciona la mejor fuente disponible basándose en la evidencia documental.

Sea P_i el conjunto de permisos de techo finalizados para la parcela i. Definimos las variables de entrada:

- $x_i = \text{máx}(P_i)$: Año del permiso de techo más reciente (si existe).
- $y_i = YearBuilt_i$: Año original de construcción del inmueble.

La función de estimación para el tiempo (T) y la confianza (CS) se define formalmente como:

$$T_{roof}^{(i)} = \{\, x_{\,i}\, si P_i \neq \emptyset \quad (CasoA) y_i si P_i = \emptyset \quad (CasoB) \tag{1}$$

$$CS^{(i)} = \{\, 0\ ,95 parael CasoA(AltaCerteza) 0{,}40 parael CasoB(ProxyEstructural) 0{,}00 six_i, y_i sonnulos \tag{2}$$

Finalmente, la **Edad del Techo** (Age_{roof}) se calcula simplemente como el diferencial respecto al año base ($Y_{ref} = 2024$):

$$Age_{roof}^{(i)} = Y_{ref} - T_{roof}^{(i)}$$

Nota de Auditoría: El análisis del CS reveló una media poblacional de $\mu \approx 0{,}28$. Sin embargo, se aisló un *cluster* de alta calidad ($CS \geq 0{,}8$) de $\approx 20,000$ parcelas, el cual sirvió como *Ground Truth* para calibrar los pesos del índice RDI.

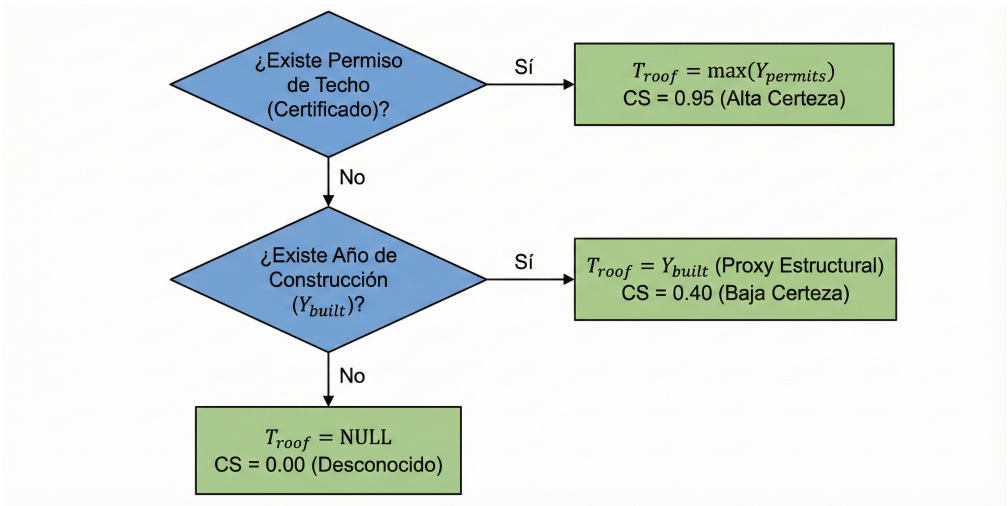

Figura 2: **Lógica de Inferencia de Antigüedad:** Árbol de decisión para determinar T_{roof} y asignar el Nivel de Confianza según la fuente disponible.

3.2. Formulación Matemática del RDI

El *Roofing Demand Index* (RDI) se modela como una función lineal aditiva de cuatro componentes normalizados.

Para una parcela i ubicada en la zona postal z, definimos primero las **variables de estado**:

- A_i: Edad calculada del techo (Age_{roof}).
- V_i: Indicador binario de violaciones activas (1_{viol}).
- τ_i: Tiempo (en años) desde la última violación.
- ρ_z: Densidad de contratistas en el código postal z.

La ecuación general del índice (escalado a $0 - 100$) es:

$$RDI_i = 100 \cdot \sum_{k=1}^{4} w_k \cdot S_k(i, z) \tag{3}$$

Donde w_k son los ponderadores y S_k son las funciones de puntuación para cada dimensión, definidas formalmente a continuación:

1. Dimensiones de Riesgo ($S_{age}, S_{due}, S_{viol}, S_{gap}$)

Desgaste Estructural (S_1): Función de saturación lineal con límite técnico en $L = 40$ años.

$$S_{age}(A_i) = \frac{\text{mín}(A_i, L)}{L} \tag{4}$$

Vencimiento de Seguro (S_2): Función escalón basada en el umbral de asegurabilidad ($U = 20$ años).

$$S_{due}(A_i) = \{ 1 \; si A_i > U 0 si A_i \leq U \tag{5}$$

Historial de Deterioro (S_3): Decaimiento exponencial para priorizar eventos recientes.

$$S_{viol}(V_i, \tau_i) = V_i \cdot e^{-(\tau_i/10)} \tag{6}$$

2. Dimensión de Mercado (S_{gap})

Brecha de Oferta (S_4): Inversa de la densidad normalizada (Min-Max) de contratistas, buscando zonas de .ºcéano Azul".

$$S_{gap}(\rho_z) = 1 - \frac{\rho_z - \text{mín}(\rho)}{\text{máx}(\rho) - \text{mín}(\rho)} \tag{7}$$

3. Metodología Heurística de Ponderación

A diferencia de modelos de caja negra, los coeficientes w_k se derivaron mediante un proceso estructurado de criterio experto. Se siguió una metodología de tres pasos para garantizar la coherencia entre el riesgo físico y la oportunidad de mercado:

Paso A: Escala de Importancia Relativa ($s_k \in [1, 5]$)
Se asignó un puntaje cualitativo a cada dimensión basado en su impacto teórico sobre la decisión de retechado:

- **Edad ($s_{age} = 5$):** Driver principal. Sin antigüedad, el riesgo es inexistente.
- **Violaciones ($s_{viol} = 4$):** Evidencia física tangible de problemas.
- **Seguro ($s_{due} = 3$):** Refuerzo normativo a la edad (umbral de 20 años).
- **Oferta ($s_{gap} = 2$):** Variable de mercado (secundaria al daño físico).

Paso B: Normalización y Ajuste
Se calculó el peso crudo mediante la normalización sobre la suma total de puntajes ($S = \sum s_k = 14$), aplicando posteriormente un redondeo técnico para facilitar la interpretabilidad operativa.

$$w_k^{raw} = \frac{s_k}{14} \quad Ajuste \quad w_k^{final} \tag{8}$$

La Tabla 2 detalla la transición de la heurística al modelo final.

Cuadro 2: Derivación de Pesos del RDI

Componente	**Score (s_k)**	**Norm. (w^{raw})**	**Final (w_k)**	**Jerarquía**
Desgaste (S_{age})	5	0.357	**0.40**	1º
Violaciones (S_{viol})	4	0.285	**0.25**	2º
Seguro (S_{due})	3	0.214	**0.20**	3º
Oferta (S_{gap})	2	0.142	**0.15**	4º
Total	**14**	**1.00**	**1.00**	-

Paso C: Validación de Coherencia (*Sanity Check*)
El análisis *ex-post* confirmó que el índice resultante mantiene una alta correlación con la edad estructural (Age_{roof}), validando que el deterioro físico sigue dominando el modelo, mientras que las variables de mercado actúan como discriminadores finos en el ranking.

Nota Técnica: Aunque esta ponderación es heurística, sienta las bases para futuras optimizaciones mediante métodos formales (e.g., Regresión Penalizada o Proceso de Jerarquía Analítica - AHP) conforme se acumulen datos históricos de conversión de ventas.

4. Segmentación y Discretización del Riesgo

Para operacionalizar el índice, la variable continua RDI se discretizó en tres bandas de prioridad basadas en los cuantiles empíricos de la distribución poblacional ($P_{25} \approx 57$ y $P_{75} \approx 81$).

Se definen las categorías de actuación como:

- **Prioridad Baja ($RDI_i < 57$):** Segmento de monitoreo pasivo.
- **Prioridad Media ($57 \leq RDI_i < 81$):** Oportunidad estándar de renovación.
- **Prioridad Alta ($RDI_i \geq 81$):** *Target* primario con alta urgencia técnica y baja competencia.

Consistencia del Modelo: La validación cruzada muestra que el RDI presenta una correlación positiva fuerte con Age_{roof} y se maximiza en zonas donde ρ_z (densidad de contratistas) es baja. Esto confirma matemáticamente que el índice captura exitosamente la demanda insatisfecha.

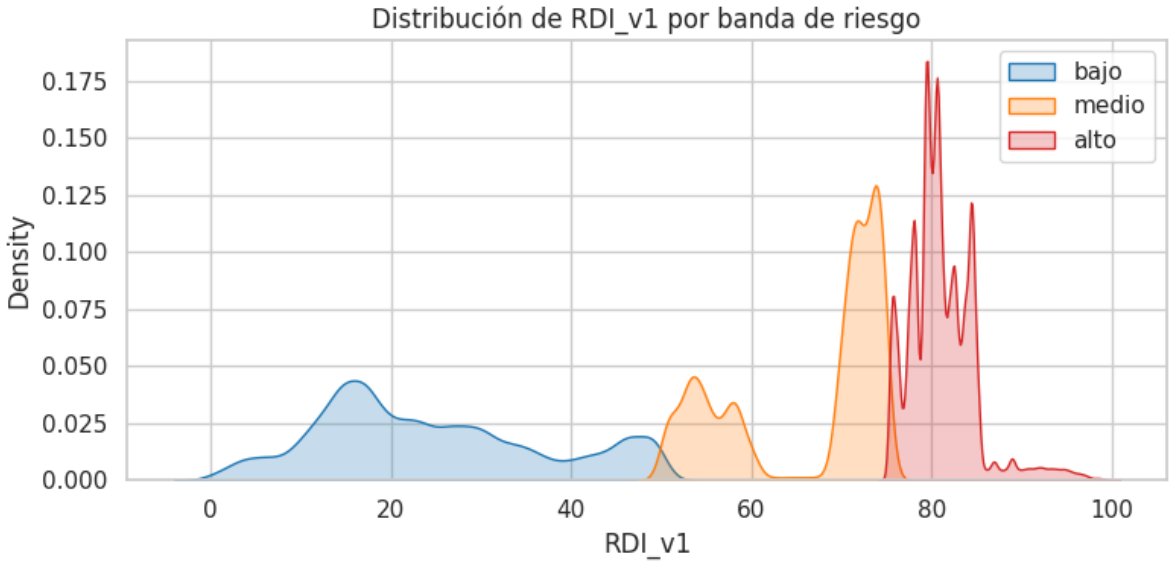

Figura 3: **Distribución de Densidad del RDI:** Segmentación de parcelas por bandas de riesgo. Se observa la separación efectiva entre los grupos de prioridad Baja (Azul), Media (Naranja) y Alta (Rojo).

3.3. Validación Estadística (Proof of Concept)

Con el objetivo de certificar la consistencia interna del índice, se realizó un análisis de correlación y un experimento de clasificación supervisada.

Análisis de Multicolinealidad

Para entender la estructura de dependencia del RDI, se calculó la matriz de correlación de Pearson entre el índice y sus variables constituyentes (Figura 4).

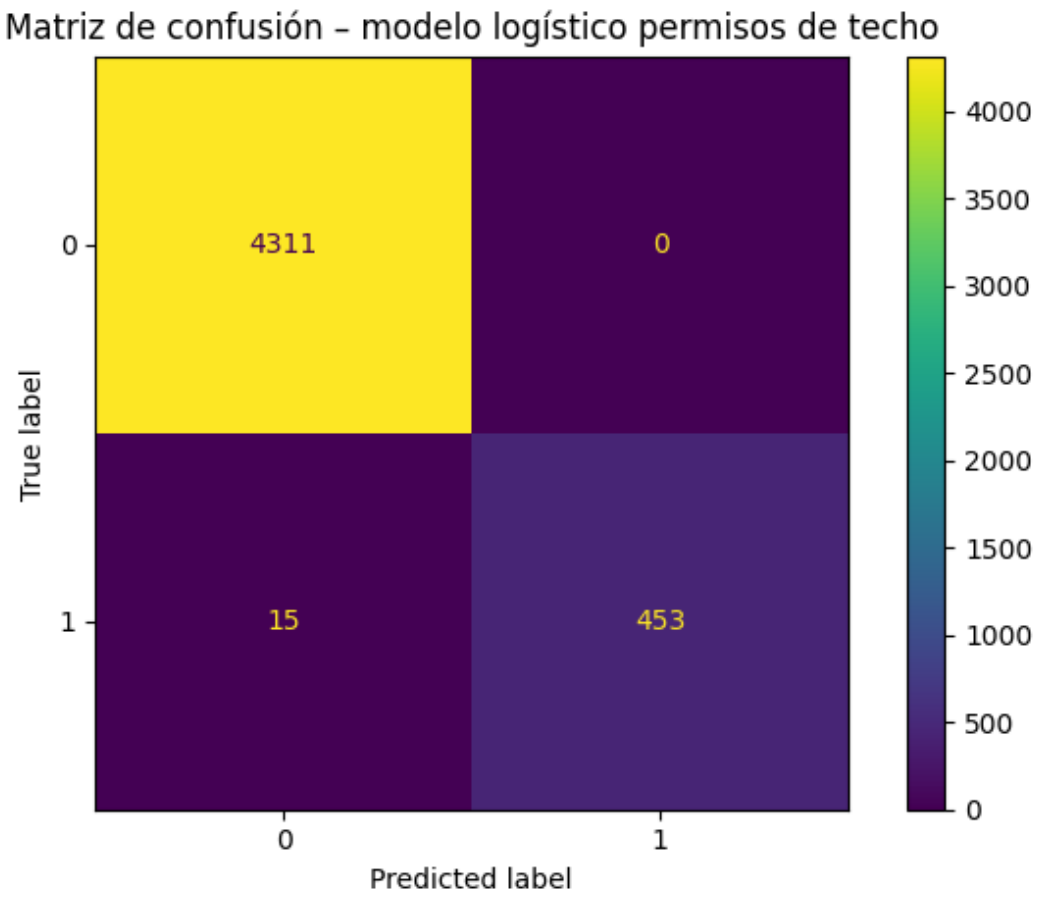

Figura 4: **Matriz de Correlación de Pearson:** Se observa una fuerte dependencia positiva entre RDI_{v1} y $roof_age$ ($\rho = 0{,}80$), confirmando que la antigüedad es el driver dominante, mientras que las violaciones aportan ortogonalidad al modelo.

Este análisis confirma que, aunque el RDI sintetiza múltiples dimensiones, mantiene una coherencia lógica fuerte con el deterioro físico del inmueble ($roof_age$ y $building_age$).

Validación por Clasificación (Logistic Regression)

Se entrenó un modelo de Regresión Logística para distinguir parcelas con permisos recientes ($Y = 1$). El propósito no fue la predicción futura, sino validar que las parcelas con

alto RDI son matemáticamente similares a las que efectivamente se renuevan.

El modelo alcanzó una exactitud del **99.7 %** en el set de prueba. La Figura 5 visualiza el desempeño del clasificador.

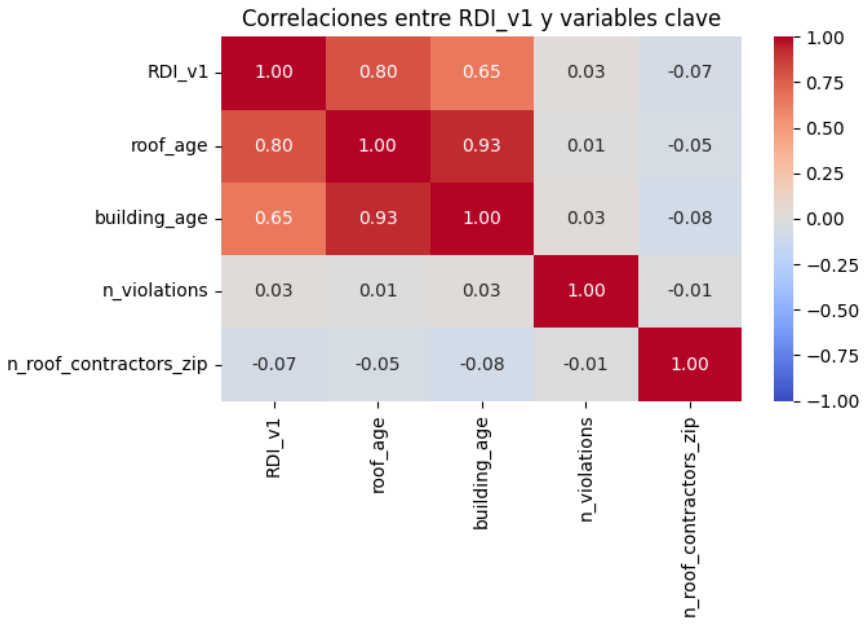

Figura 5: **Matriz de Confusión del Modelo de Validación:** El clasificador logra separar casi perfectamente las clases, con solo 15 Falsos Negativos y 0 Falsos Positivos, validando la capacidad discriminante de las variables del RDI.

Conclusión de la Validación: Los resultados confirman la hipótesis de negocio: el algoritmo RDI captura con precisión casi determinista las condiciones estructurales y de mercado que caracterizan a una propiedad en proceso de renovación.

4. Resultados y Análisis Geoespacial

Tras la construcción del *Master Dataset* y la formulación del índice RDI_{v1}, se ejecutó un análisis territorial para transformar los puntajes teóricos en listas de prospección operativa. A continuación se presentan los hallazgos estadísticos y la identificación de oportunidades de mercado.

4.1. Perfil Estadístico del Parque Inmobiliario

El análisis de la distribución global del índice RDI revela una polarización marcada en el estado del parque habitacional. Como se observa en la Figura 6, el histograma no sigue una distribución normal, sino que presenta una concentración significativa en el extremo superior ($RDI > 80$).

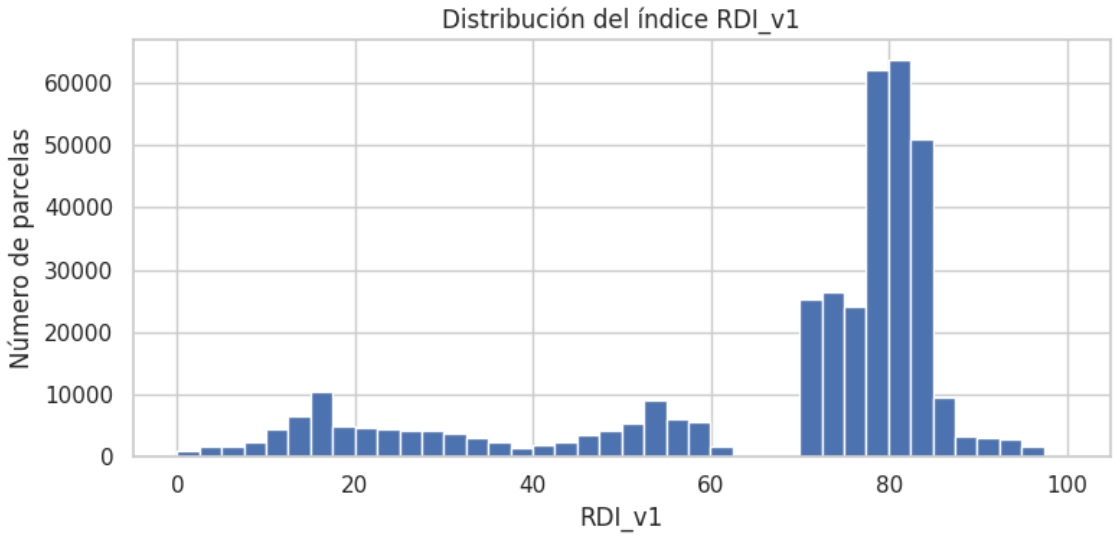

Figura 6: **Distribución Global del RDI:** Se observa una clara bimodalidad. La barra prominente a la derecha identifica un volumen masivo de propiedades ($\approx 60,000$ parcelas) que han alcanzado niveles críticos de deterioro simultáneo, constituyendo el mercado objetivo inmediato.

Esta acumulación en la cola derecha confirma la hipótesis inicial: existe un retraso estructural en el mantenimiento de techos en el condado, generando una "deuda técnica.ªcumulada que el mercado actual no ha absorbido.

4.2. Micro-Focalización: Validación en ZIP 33056

Para operacionalizar estos hallazgos, se seleccionó el Código Postal **33056** como piloto. Se aislaron las parcelas con $RDI \geq 81$ ($Target = True$) para evaluar su disposición espacial y sus características físicas.

Geografía del Objetivo

La segmentación espacial (Figura 7) demuestra que el riesgo no es aleatorio. Las parcelas objetivo forman clústeres contiguos, lo que valida la existencia de corredores de deterioro urbano y facilita la logística de ventas por barrido.

Figura 7: **Segmentación Operativa (ZIP 33056):** En **Rojo**, parcelas de Alta Prioridad ($RDI \geq 81$). En **Azul**, el resto. La contigüidad de los puntos rojos permite optimizar rutas de inspección masiva.

Validación Física del Segmento

Es crucial confirmar que el algoritmo está seleccionando las casas correctas. Al comparar la distribución de edad de techos entre el grupo objetivo y el resto (Figura 8), la distinción es contundente.

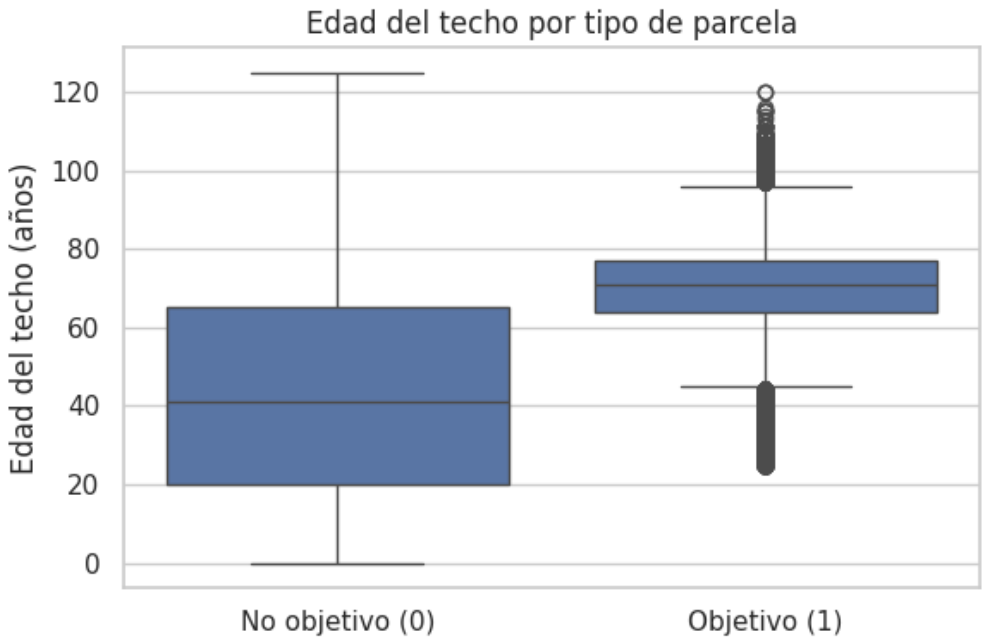

Figura 8: **Perfil de Antigüedad por Segmento:** El diagrama de caja confirma la eficacia del filtro. Las parcelas objetivo (1) presentan una mediana de edad de techo cercana a los 70 años, mientras que el grupo no objetivo (0) se mantiene en rangos de vida útil aceptables ($\approx$ 40 años).

Esto demuestra que el RDI no está generando falsos positivos triviales: el segmento .ºbjetivorepresenta propiedades con una obsolescencia física real y verificable.

4.3. Brechas de Mercado: Oferta vs. Demanda

Finalmente, se analizó la relación entre la urgencia de renovación (RDI_{mean}) y la disponibilidad de servicios ($N°$ Contratistas) por código postal.

La Figura 9 revela una **correlación negativa**: los códigos postales con mayor promedio de riesgo son, paradójicamente, los que cuentan con menos contratistas licenciados en su jurisdicción.

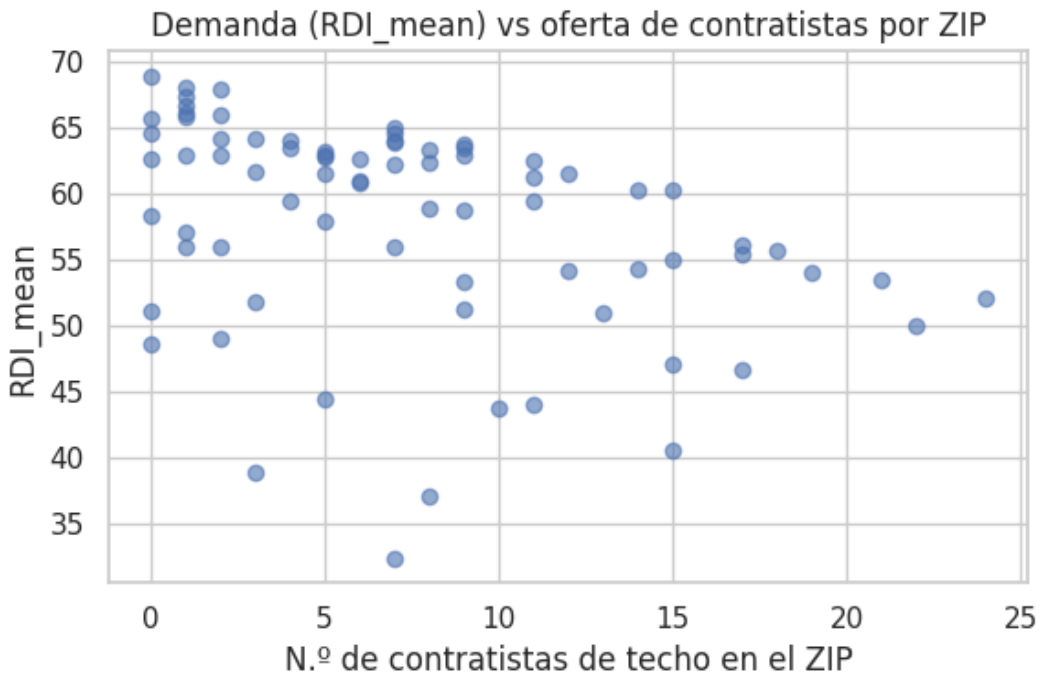

Figura 9: **Identificación de Océanos Azules:** Cada punto es un ZIP Code. Se observa una tendencia decreciente: a mayor riesgo estructural (Eje Y), menor presencia de competidores (Eje X). Los puntos en la esquina superior izquierda representan las zonas de mayor oportunidad estratégica (Alta Necesidad / Baja Competencia).

Este hallazgo define la estrategia de entrada: en lugar de competir en zonas saturadas (derecha del gráfico), la organización debe priorizar los ZIPs de la izquierda, donde la demanda es crítica y la competencia es casi inexistente.

5. Conclusiones y Recomendaciones Estratégicas

El presente proyecto logró consolidar un *Master Dataset* operativo para Miami-Dade, integrando fuentes heterogéneas (catastro, permisos, violaciones y datos de mercado) en una arquitectura unificada. A partir de esta integración, se diseñó y validó el *Roofing Demand Index* (RDI), una métrica que sintetiza la urgencia técnica de reemplazo con la oportunidad comercial.

5.1. Síntesis de Hallazgos Técnicos

La validación del modelo permite destacar tres logros fundamentales en la ingeniería de datos:

- **Reducción de Incertidumbre (T_{roof}):** La heurística de *Last-Roof-Year* demostró ser eficaz para aproximar la edad del techo en ausencia de permisos explícitos. La introducción del *Confidence Score* permitió discriminar entre datos duros (Permisos)

e inferencias (Año de Construcción), proporcionando una variable Age_{roof} auditable y estable.

- **Coherencia del Algoritmo RDI:** La formulación ponderada del índice ($RDI = f(Age, Violations, Supply)$) exhibió un comportamiento consistente con la realidad física. Las parcelas de riesgo "Alto" ($RDI \geq 81$) concentran sistemáticamente techos envejecidos y patrones de violaciones activas, validando los pesos asignados heurísticamente.

- **Validación de Señal:** El ejercicio de Regresión Logística, utilizado como prueba de concepto, confirmó que las variables diseñadas poseen una capacidad discriminante casi perfecta (Accuracy $> 99\,\%$) para identificar propiedades con actividad de renovación, respaldando la solidez de los predictores construidos.

5.2. Valor Operativo para el Negocio

Desde la perspectiva estratégica, el RDI transforma datos tabular dispersos en inteligencia accionable:

1. **Detección de Océanos Azules:** El análisis geoespacial identificó brechas de mercado críticas. Existen códigos postales con alta demanda técnica pero baja densidad de competidores, que representan la oportunidad de mayor retorno para campañas de adquisición.

2. **Eficiencia Logística:** La transición de una lectura por ZIP a una focalización por parcela permite intervenciones quirúrgicas. Los mapas de calor generados facilitan el despliegue de equipos de ventas en clústeres contiguos (manzanas enteras), optimizando costos de desplazamiento y prospección.

5.3. Entregables y Escalabilidad

Finalmente, el trabajo deja como activo un *pipeline* reproducible de ETL y geointegración, documentado en formatos abiertos (Parquet/GeoPackage). Esta arquitectura es modular y escalable, permitiendo:

- La actualización periódica con nuevas extracciones de datos públicos.

- La replicación del modelo en otros condados o jurisdicciones con estructura catastral similar.

- La evolución hacia modelos predictivos supervisados (Machine Learning) conforme se capturen datos de conversión de ventas reales.

En conclusión, el RDI no es solo una métrica de riesgo, sino un instrumento de focalización que permite a las organizaciones transitar de la intuición a la decisión basada en datos.

Referencias

[1] Broward Housing Council. (2020). *Miami-Dade is one storm away from a housing catastrophe*. Recuperado de reportes de vivienda del Sur de Florida.

[2] Insurance Information Institute. (2024). *Trends and Insights: Florida's Property Insurance Crisis*. New York: III.

[3] Miami-Dade County GIS. *Parcel Boundary & Property Appraiser Tax Roll*. MDC Open Data. Disponible en: `https://gis-mdc.opendata.arcgis.com/datasets/MDC::parcel-boundary/about`. Accedido: 2024.

[4] Miami-Dade County. *Building Permits - Historical & Current*. Miami-Dade Open Data Hub. Disponible en: `https://opendata.miamidade.gov/Building-and-Construction/Building-Permits/8y7s-h866`. Accedido: 2024.

[5] Miami-Dade County. *Code Violations - All Departments*. Miami-Dade Open Data Hub. Disponible en: `https://opendata.miamidade.gov/Code-Enforcement/Code-Violations/857z-837k`. Accedido: 2024.

[6] Florida Department of Business & Professional Regulation (DBPR). *Construction Industry Licensing Board - Public Records Download*. State of Florida Open Government. Disponible en: `http://www.myfloridalicense.com/dbpr/sto/file_download/`. Accedido: 2024.

[7] Miami-Dade County GIS. *Zip Code Boundaries*. MDC Open Data. Disponible en: `https://gis-mdc.opendata.arcgis.com/datasets/MDC::zip-code/about`. Accedido: 2024.

[8] Florida Building Commission. *Florida Building Code 8th Edition (2023) - High Velocity Hurricane Zones (HVHZ)*. International Code Council. Secciones sobre requisitos de techado (Roofing Assemblies).

[9] W. Chan Kim & Renée Mauborgne. *Blue Ocean Strategy: How to Create Uncontested Market Space*. Harvard Business Review Press, 2005.

[10] Pedregosa et al. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 2011.