



Pontificia Universidad
JAVERIANA
Cali

**MODELO ANALÍTICO PARA LA PREDICCIÓN DE GARANTÍAS POSVENTA EN PROYECTOS
RESIDENCIALES: APLICACIÓN EN UNA EMPRESA CONSTRUCTORA COLOMBIANA**

DIANA CAMILA CASTILLO NARVÁEZ

MAICOL STIVEN CONTRERAS MACÍAS

*Proyecto Aplicado para optar al título de Magister en Ciencia de
Datos*

Director

MsG David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, MAYO 2026

FICHA RESUMEN

MODELO ANALÍTICO PARA LA PREDICCIÓN DE GARANTÍAS POSVENTA EN PROYECTOS RESIDENCIALES: APLICACIÓN EN UNA EMPRESA CONSTRUCTORA COLOMBIANA

1. ÁREA DE TRABAJO: Construcción
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTES:
 - Diana Camila Castillo Narváez
 - Maicol Contreras Macías
4. DIRECTOR: David Arango Londoño
5. VINCULACIÓN DEL DIRECTOR: Docente Universidad Javeriana Cali
6. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co
7. GRUPO O EMPRESA QUE LO AVALA: Constructora anónima
8. PALABRAS CLAVE: Aprendizaje automático, reclamaciones posventa, modelo predictivo, construcción de vivienda, optimización operativa
9. FECHA DE INICIO: 21 julio 2025
10. DURACIÓN ESTIMADA : 10 meses
11. RESUMEN:

Este proyecto aplicado tuvo como objetivo desarrollar un modelo predictivo basado en técnicas de aprendizaje automático para estimar el comportamiento de las reclamaciones por garantías posventa en proyectos de vivienda entregados por una constructora colombiana. A partir del análisis de bases de datos históricas del sistema CRM, información de entregas de inmuebles, características técnicas de los proyectos, tipo de acabado y variables contextuales exploratorias, se buscó estimar la cantidad esperada, la categoría técnica y el momento de ocurrencia de las reclamaciones. El trabajo incluyó tareas de limpieza, integración y transformación de datos, homologación de proyectos y categorías de daño, construcción de la exposición mensual de inmuebles, ingeniería de características, selección de variables relevantes y evaluación de modelos supervisados de regresión. El modelo seleccionado fue XGBoost con variables de historial reciente, debido a que presentó el mejor equilibrio general entre las métricas evaluadas, incluyendo MAE, RMSE, R^2 , MAPE sobre observaciones positivas y comparación entre totales reales y predichos. Como resultado, el modelo permitió generar predicciones futuras para apoyar la planeación operativa del área de posventa; para 2026 se estimó una carga aproximada de 4.247 locativas, con análisis por mes, proyecto y padre/capítulo técnico. En un sector como el de la construcción, con una gestión posventa tradicionalmente reactiva y con oportunidades de fortalecimiento en el uso analítico de los datos, esta propuesta representa una aplicación práctica de ciencia de datos para anticipar cargas de atención, priorizar recursos, identificar categorías críticas y apoyar la toma de decisiones basada en evidencia. El proyecto contribuye al fortalecimiento de la gestión proactiva de garantías y a la transformación digital del proceso posventa, dejando como producto un modelo predictivo reproducible y resultados estructurados para su consulta mediante una herramienta de visualización.

TABLA DE CONTENIDO

INTRODUCCIÓN	7
1. CONTEXTUALIZACIÓN DEL PROYECTO	9
1.1. Definición del problema	9
1.1.1. Planteamiento del problema	9
1.1.2. Formulación del problema	11
1.1.3. Sistematización	11
1.2. Objetivos	12
1.2.1. Objetivo general	12
1.2.2. Objetivos específicos	12
1.3. Alcance del proyecto	12
2. MARCO TEORICO Y ANTECEDENTES	14
2.1. Garantías posventa en proyectos de vivienda	14
2.2. Analítica predictiva y aprendizaje automático supervisado	14
2.3. Modelos de regresión para datos de conteo	15
2.4. Modelos basados en árboles y XGBoost	16
2.5. Calidad de datos, preparación e ingeniería de características	17
2.6. Metodología CRISP-DM	17
2.7. Antecedentes	18
3. PREPARACIÓN Y ESTRUCTURACIÓN DE LOS DATOS	20
3.1. Fuentes de datos utilizadas	20
3.2. Limpieza, depuración y homologación	25
3.3. Integración de entregas, garantías y características del proyecto	27
3.4. Construcción de la exposición de inmuebles	30
3.5. Definición de la variable objetivo	30
3.6. Construcción del dataset final de modelado	31
3.7. Consideraciones metodológicas y exclusiones	32
3.8. Caracterización del dataset final de modelado	35
4. MODELADO PREDICTIVO SUPERVISADO	36
4.1. Enfoque general del modelado	36

4.2.	Modelos evaluados	37
4.3.	Estrategia de entrenamiento y validación	38
4.4.	Evaluación cuantitativa de desempeño	39
4.5.	Selección del modelo final	41
4.6.	Interpretación de variables del modelo.....	43
5.	RESULTADOS Y PREDICCIÓN FUTURA	48
5.1.	Justificación del horizonte de aplicación futura	48
5.2.	Aplicación del modelo sobre entregas futuras	49
5.3.	Predicción mensual de locativas para 2026.....	49
5.4.	Resultados por proyecto	51
5.5.	Resultados por padre o capítulo	52
5.6.	Lectura conjunta: mes, proyecto y padre/capítulo	54
5.7.	Implicaciones para la gestión posventa	54
5.8.	Discusión de resultados	55
5.9.	Alcance de los resultados del modelo	56
6.	HERRAMIENTA WEB DE VISUALIZACIÓN	57
6.1.	Propósito de la herramienta	57
6.2.	Archivo de entrada y estructura de resultados	58
6.3.	Vistas principales de la página web	59
6.4.	Filtros e interactividad	62
6.5.	Uso operativo de la herramienta	62
6.6.	Valor aplicado para la gestión posventa	63
6.7.	Ejemplo de consulta operativa	64
7.	CONCLUSIONES Y TRABAJOS FUTUROS	65
7.1.	Limitaciones del proyecto	66
7.2.	Trabajos futuros	67
	BIBLIOGRAFÍA	69

LISTA DE FIGURAS

Ilustración 1. Flujo general de preparación y estructuración de datos	34
Ilustración 2. Comparación de modelos evaluados	41
Ilustración 3. Importancia agrupada de variables	47
Ilustración 4. Predicción mensual de locativas para 2026	51
Ilustración 5. Proyectos con mayor predicción de locativas en 2026	52
Ilustración 6. Padres/capítulos con mayor predicción de locativas en 2026	54

LISTA DE TABLAS

Tabla 1. Fuentes de datos utilizadas	24
Tabla 2. Criterios de depuración y exclusión metodológica	26
Tabla 3. Variables finales utilizadas en el dataset de modelado	28
Tabla 4. Construcción de la variable objetivo y unidad de análisis	31
Tabla 5. Variables climáticas exploradas y decisión metodológica	34
Tabla 6. Características del dataset final de modelado	35
Tabla 7. Comparación de modelos evaluados	40
Tabla 8. Métricas del modelo seleccionado.....	42
Tabla 9. Hiperparámetros del modelo XGBoost seleccionado	43
Tabla 10. Importancia agrupada de variables del modelo seleccionado	45
Tabla 11. Predicción mensual de locativas para 2026	50
Tabla 12. Proyectos con mayor predicción de locativas en 2026.....	51
Tabla 13. Padres/capítulos con mayor predicción de locativas en 2026.....	53

INTRODUCCIÓN

La gestión de garantías posventa constituye una etapa crítica dentro del ciclo de vida de los proyectos residenciales, debido a que incide directamente en la experiencia del cliente, la percepción de calidad del inmueble y la eficiencia operativa de la constructora. En esta fase se reciben, clasifican y atienden reclamaciones asociadas a fallas, daños o inconformidades reportadas después de la entrega de las viviendas, lo que exige coordinar recursos técnicos, contratistas, materiales, agendas de atención y procesos de seguimiento. Sin embargo, en muchas organizaciones del sector construcción, la gestión posventa continúa abordándose principalmente desde una lógica reactiva, es decir, una vez el cliente ha reportado la reclamación.

La constructora en estudio cuenta con información histórica relevante sobre entregas de inmuebles, reclamaciones de garantías, características de proyectos, categorías técnicas de daño y condiciones de entrega. No obstante, antes del desarrollo de este proyecto, dicha información no se encontraba integrada en un modelo predictivo que permitiera anticipar el comportamiento futuro de las garantías posventa. Esta situación limitaba la posibilidad de estimar cargas operativas, identificar proyectos o categorías con mayor probabilidad de reclamación y planear recursos de manera anticipada.

En este contexto, el presente proyecto aplicado desarrolló un modelo predictivo basado en técnicas de aprendizaje automático supervisado para estimar la cantidad esperada de locativas de garantía en proyectos residenciales. Para ello, se realizó un proceso de preparación e integración de datos históricos de entregas, locativas de garantía, homologación de macroproyectos, estandarización de padres o capítulos técnicos, incorporación de características estructurales de los proyectos y clasificación del tipo de acabado. Adicionalmente, se construyó una estructura de exposición mensual de inmuebles entre los meses 0 y 12 posteriores a la entrega, permitiendo que el modelo aprendiera tanto de los casos donde se presentaron reclamaciones como de los escenarios donde existían inmuebles expuestos sin reportes registrados.

El problema fue formulado como una tarea de regresión supervisada para datos de conteo, en la cual la variable objetivo correspondió al número de locativas reales observadas para combinaciones de periodo, macroproyecto, mes desde entrega y padre/capítulo homologado. Durante el proceso de modelado se evaluaron diferentes alternativas, incluyendo una línea base de tasa histórica, Random Forest, regresión de Poisson, XGBoost Poisson y XGBoost con variables de historial reciente. El modelo seleccionado fue XGBoost con historial, debido a que presentó el mejor equilibrio general entre las métricas evaluadas.

Posteriormente, el modelo seleccionado fue aplicado sobre las entregas futuras proyectadas, con el fin de estimar la carga esperada de garantías para 2026. Los resultados obtenidos permiten generar una lectura anticipada de posibles cargas de atención y ofrecen insumos para la planeación de recursos, la priorización de seguimiento y la preparación de acciones preventivas.

El proyecto también dejó estructurados los resultados predictivos para su consulta mediante una herramienta web de visualización, orientada a facilitar la exploración de las predicciones por mes, proyecto y padre/capítulo. Esta herramienta se plantea como un puente entre el modelo analítico y el uso operativo por parte del negocio, permitiendo que los resultados sean interpretados de forma más accesible por usuarios técnicos y directivos.

El documento se organiza en siete capítulos. El primero presenta la contextualización del proyecto, el problema, los objetivos y el alcance. El segundo desarrolla el marco teórico y los antecedentes relevantes. El tercero describe la preparación y estructuración de los datos. El cuarto presenta el proceso de modelado predictivo, la evaluación de modelos y la selección del modelo final. El quinto expone los resultados y la predicción futura para 2026. El sexto presenta la herramienta web de visualización desarrollada para consultar los resultados predictivos. Finalmente, el séptimo capítulo reúne las conclusiones, limitaciones y trabajos futuros.

En conjunto, este trabajo evidencia el potencial de la ciencia de datos para transformar registros históricos de garantías en información útil para la toma de decisiones. La solución desarrollada no reemplaza el criterio técnico del área de posventa, pero sí aporta una base cuantitativa para anticipar cargas, identificar focos de seguimiento y avanzar hacia una gestión más proactiva, trazable y fundamentada en evidencia dentro del sector constructor.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. Definición del problema

El sector de la construcción ha presentado limitaciones significativas en la adopción de herramientas analíticas avanzadas que permitan transformar datos operativos en decisiones estratégicas. En la etapa de posventa, estas limitaciones se reflejaban en una gestión mayoritariamente reactiva: las constructoras atendían reclamaciones solo después de que los clientes las reportaban, sin mecanismos formales para anticipar fallas, identificar patrones o estimar comportamientos futuros.

En la constructora del estudio existía una base de datos extensa y detallada sobre reclamaciones posventa, consolidada a través del sistema CRM. No obstante, esta información no había sido utilizada para desarrollar modelos predictivos que permitieran anticipar la cantidad, tipo o distribución temporal de las reclamaciones. Como consecuencia, decisiones clave como la programación de visitas técnicas, la asignación de recursos, la estimación de cargas operativas o la identificación de defectos recurrentes se realizaban sin soporte analítico.

El problema central radicaba, por tanto, en la ausencia de un sistema basado en ciencia de datos que transformara los registros históricos en conocimiento útil para la gestión proactiva. La falta de un modelo predictivo limitaba la posibilidad de planificar de manera anticipada, optimizar recursos, reducir reprocesos y mejorar la experiencia del cliente durante la etapa posventa, considerada crítica en la percepción final de calidad del inmueble y de la marca constructora.

1.1.1. Planteamiento del problema

El sector de la construcción presentó durante los últimos años un rezago estructural en términos de adopción tecnológica. De acuerdo con el McKinsey & Company group [1], esta industria se ubicó entre las menos digitalizadas a nivel mundial, ocupando los últimos lugares en el índice de digitalización. Esta situación limitó de manera significativa la capacidad de las empresas para aprovechar sus datos, optimizar procesos, anticipar problemas y mejorar la toma de decisiones operativas.

Una de las áreas más afectadas por este rezago fue la etapa de posventa, un componente esencial de la experiencia del cliente y del ciclo de vida del proyecto. A pesar de ser un proceso regulado por la normativa colombiana y contar con protocolos establecidos, la gestión posventa se continuó desarrollando con un enfoque predominantemente reactivo: las reclamaciones se atendían únicamente cuando los clientes las reportaban, sin herramientas que permitieran anticipar fallas, identificar tendencias o planificar intervenciones preventivas.

En la constructora esta situación se hizo evidente. Aunque la empresa disponía de registros históricos detallados de miles de reclamaciones generadas en proyectos anteriores — incluyendo fechas de entrega, tipologías constructivas, clasificación de daños, tiempos de atención, espacios afectados y características del inmueble— dicha información no se había utilizado para desarrollar modelos predictivos que permitieran anticipar eventos o comprender el comportamiento posventa de forma sistemática. Las decisiones operativas, como la asignación de cuadrillas, la priorización de casos, la programación de visitas técnicas o la estimación de cargas de trabajo, continuaban basándose en la experiencia del personal y en observaciones directas, y no en análisis cuantitativos derivados de los datos.

La ausencia de un sistema analítico impedía transformar el conocimiento almacenado en el CRM en información útil para la planificación estratégica. Por ejemplo, variables como el tipo de proyecto, la tipología del daño, la estacionalidad climática, la clasificación del inmueble o el tiempo transcurrido desde la entrega podían aportar señales de alerta relevantes, pero no estaban siendo utilizadas para identificar patrones, calcular probabilidades o anticipar comportamientos futuros. Esto generaba riesgos como la subestimación de cargas de trabajo, la falta de preparación ante picos de reclamos, la repetición de fallas ya conocidas y la pérdida de oportunidades para mejorar la calidad constructiva.

Asimismo, la carencia de modelos predictivos dificultaba la identificación de relaciones entre variables internas y externas. Factores como condiciones climáticas, características del diseño arquitectónico, tipo de materiales utilizados, historial del proyecto, procesos constructivos o incluso la gestión del subcontratista podían estar asociados a ciertos tipos de reclamaciones. Sin embargo, sin un análisis sistemático era imposible estimar su impacto o incorporarlos en la toma de decisiones.

Otro elemento crítico fue la imposibilidad de anticipar cómo evolucionaban los reclamos a lo largo del tiempo. La distribución temporal de las reclamaciones —especialmente en los primeros 12 meses posteriores a la entrega— contiene patrones valiosos para comprender el comportamiento de los proyectos y detectar anomalías. No obstante, sin un modelo de predicción que integrara información histórica, la constructora no contaba con una herramienta que permitiera estimar en qué momento era más probable que se presentaran determinadas fallas o qué tipos de reclamaciones podían esperarse según la tipología de vivienda.

En consecuencia, la gestión operativa se veía limitada, los recursos se asignaban de manera reactiva y las estrategias para mejorar la experiencia del cliente carecían de soporte analítico. Además, la empresa perdía la oportunidad de retroalimentar procesos constructivos con evidencia derivada de los datos, lo cual es especialmente relevante en una industria donde la repetición de errores genera sobrecostos significativos.

Este panorama evidenció la necesidad de desarrollar un modelo predictivo capaz de estimar el tipo, la cantidad y la distribución temporal de las reclamaciones posventa, integrando datos históricos multivariados y técnicas de aprendizaje automático. Un modelo de esta naturaleza

permitiría evolucionar hacia una gestión preventiva, fortalecer la toma de decisiones basadas en evidencia, identificar variables críticas, anticipar riesgos operativos y optimizar la asignación de recursos. Además, facilitaría la generación de alertas tempranas, la priorización de intervenciones y el diseño de estrategias orientadas a mejorar la calidad constructiva, contribuyendo a la modernización y transformación digital del proceso posventa en la constructora del análisis.

1.1.2. Formulación del problema

A pesar de contar con bases de datos históricas extensas sobre las reclamaciones posventa registradas en el sistema CRM, la constructora no disponía de un modelo predictivo que permitiera estimar el comportamiento de dichas reclamaciones en términos de tipo, cantidad y distribución temporal. La falta de un sistema analítico impedía anticipar fallas, planificar recursos y adoptar un enfoque preventivo en la gestión posventa. En este contexto, surgió la necesidad de responder la siguiente pregunta central de investigación:

Pregunta de investigación:

¿Cómo estimar el comportamiento posventa de proyectos de vivienda de la constructora del presente estudio en términos de tipo, volumen y distribución temporal de las reclamaciones, mediante el entrenamiento de modelos supervisados sobre datos históricos multivariados que integran variables estructurales, temporales, operativas y contextuales?

1.1.3. Sistematización

Para abordar la formulación general, se definieron las siguientes preguntas específicas que guiaron el desarrollo del proyecto:

- ¿Qué variables del histórico de reclamaciones posventa tuvieron mayor relevancia para estimar la cantidad esperada de garantías posventa por proyecto, padre/capítulo técnico y mes transcurrido desde la entrega del inmueble?
- ¿Qué técnicas de aprendizaje automático resultaron más adecuadas para modelar el comportamiento de las reclamaciones posventa en función de sus características temporales y tipológicas?
- ¿Qué nivel de precisión, sensibilidad y capacidad de generalización alcanzaron los modelos predictivos al ser validados con datos reales?
- ¿Qué recomendaciones pudieron derivarse de los resultados del modelo para apoyar la toma de decisiones en la gestión proactiva de la posventa y la asignación eficiente de recursos?

1.2.Objetivos

1.2.1. Objetivo general

Desarrollar un modelo predictivo basado en técnicas de aprendizaje automático que permita estimar el comportamiento de las reclamaciones posventa en proyectos de vivienda, considerando su cantidad esperada, categoría de daño y momento de ocurrencia posterior a la entrega del inmueble, con el fin de apoyar una gestión más proactiva, anticipada y fundamentada en datos.

1.2.2. Objetivos específicos

- Preprocesar los datos históricos de reclamaciones posventa, entregas de inmuebles y características de los proyectos, mediante procesos de limpieza, integración, homologación y estructuración de la información relevante para el análisis.
- Explorar y desarrollar modelos de aprendizaje automático supervisado que permitan representar el comportamiento de las reclamaciones posventa a partir de variables asociadas al proyecto, la tipología de vivienda, el tipo de acabado, la categoría de daño, el tiempo transcurrido desde la entrega, la exposición de inmuebles y el comportamiento histórico reciente de las reclamaciones.
- Evaluar el desempeño de los modelos desarrollados mediante métricas estadísticas acordes con el problema predictivo, tales como el error absoluto medio, la raíz del error cuadrático medio, el coeficiente de determinación, el error porcentual sobre observaciones positivas y la comparación entre totales reales y predichos.
- Proponer lineamientos para la integración de los resultados del modelo predictivo en una herramienta de visualización y consulta, orientada a fortalecer la planeación de recursos, la generación de alertas tempranas, la priorización de intervenciones y la mejora de la experiencia del cliente.

1.3. Alcance del proyecto

El alcance del proyecto se concentró en la predicción agregada de locativas de garantía asociadas a inmuebles privados entregados por la constructora en el suroccidente Colombiano. El modelo no predice la ocurrencia individual de una reclamación, sino la cantidad esperada de locativas para combinaciones de periodo, macroproyecto, mes desde entrega y padre/capítulo homologado. La ventana modelable se definió entre los meses 0 y 12 posteriores a la entrega del inmueble, incluyendo el mes 12 por coherencia con la operación posventa.

Se excluyeron del modelo principal los reportes de zonas comunes, las locativas sin entrega asociada, los reportes antes de la entrega y los registros posteriores a la ventana modelable.

Estas exclusiones no implican que dichos casos carezcan de importancia operativa, sino que requieren estructuras de exposición diferentes o información adicional no disponible para esta versión del modelo.

Las variables climáticas fueron consideradas como fuentes contextuales exploratorias, especialmente por su posible relación con reclamaciones de cubiertas, humedades, filtraciones y goteras. Sin embargo, no fueron incorporadas como predictores principales en el modelo final debido a limitaciones de cobertura, completitud y consistencia temporal.

2. MARCO TEORICO Y ANTECEDENTES

2.1. Garantías posventa en proyectos de vivienda

La etapa de posventa en proyectos residenciales comprende las actividades de recepción, evaluación, gestión y atención de reclamaciones realizadas por los propietarios después de la entrega del inmueble. Estas reclamaciones pueden estar asociadas a defectos constructivos, acabados, redes técnicas, elementos arquitectónicos, humedades, filtraciones u otros aspectos que afectan la funcionalidad, habitabilidad o percepción de calidad de la vivienda.

En Colombia, la garantía legal de los bienes inmuebles se encuentra enmarcada en el Estatuto del Consumidor, Ley 1480 de 2011, y en el Decreto 735 de 2013, que reglamenta la efectividad de la garantía. Este marco normativo establece responsabilidades para los productores y proveedores frente a la calidad, idoneidad, seguridad y buen estado de los bienes entregados, incluyendo disposiciones específicas sobre bienes inmuebles [2], [3].

Desde el punto de vista operativo, las reclamaciones posventa representan una fuente de información valiosa para las constructoras, ya que permiten identificar patrones de fallas, categorías recurrentes, proyectos con mayor carga de atención y posibles oportunidades de mejora en procesos constructivos. Sin embargo, cuando esta información se utiliza únicamente de manera descriptiva o reactiva, se desaprovecha su potencial para anticipar comportamientos futuros. La literatura sobre defectos posentrega en vivienda muestra que las fallas detectadas después de la entrega pueden afectar la satisfacción del cliente y generar reprocesos para las empresas constructoras. Estudios sobre defectos en edificaciones residenciales han señalado que una parte importante de los problemas reportados durante el periodo de garantía se relaciona con defectos de acabado, instalaciones, elementos funcionales y condiciones constructivas que pudieron ser identificadas o mitigadas mediante un mejor seguimiento de calidad [4], [5].

En este sentido, la gestión de garantías posventa no debe entenderse únicamente como una obligación posterior a la entrega, sino como un mecanismo de retroalimentación para la mejora continua. Los registros históricos de reclamaciones permiten construir conocimiento sobre el comportamiento real de los proyectos una vez son habitados, lo cual puede apoyar decisiones de planeación, control técnico y priorización operativa.

2.2. Análítica predictiva y aprendizaje automático supervisado

La analítica predictiva se refiere al uso de datos históricos, técnicas estadísticas y algoritmos de aprendizaje automático para estimar comportamientos futuros o resultados esperados. A diferencia de la analítica descriptiva, que resume lo ocurrido en el pasado, la analítica predictiva busca identificar patrones que puedan ser utilizados para anticipar eventos, riesgos o necesidades futuras.

En el contexto de este proyecto, la analítica predictiva se aplicó a la estimación de locativas de garantía posventa. El objetivo no fue únicamente describir cuántas reclamaciones ocurrieron en periodos anteriores, sino utilizar los patrones históricos para estimar cuántas podrían presentarse bajo determinadas condiciones de exposición, proyecto, categoría técnica y tiempo desde la entrega. El aprendizaje automático supervisado es una rama de la ciencia de datos en la cual los modelos aprenden a relacionar variables explicativas con una variable objetivo conocida. En problemas de regresión, la variable objetivo es numérica y el modelo busca estimar su valor esperado a partir de los predictores disponibles. Para este proyecto, la variable objetivo correspondió a la cantidad de locativas reales observadas en una combinación específica de periodo, macroproyecto, mes desde entrega y padre/capítulo homologado.

La aplicación de aprendizaje automático en el sector construcción ha crecido en años recientes, especialmente en tareas relacionadas con planeación, control de costos, predicción de riesgos, productividad, mantenimiento y análisis de calidad. Revisiones recientes sobre inteligencia artificial y aprendizaje automático en el ciclo de vida de proyectos de construcción muestran que estas técnicas tienen potencial para mejorar la toma de decisiones, aunque también enfrentan retos asociados a la disponibilidad, calidad e integración de los datos [6].

2.3. Modelos de regresión para datos de conteo

La variable objetivo de este proyecto corresponde a un conteo: el número de locativas de garantía observadas para una combinación definida de periodo, proyecto, mes desde entrega y padre/capítulo técnico. Este tipo de variable toma valores enteros no negativos y representa la frecuencia de ocurrencia de un evento dentro de una unidad de análisis específica. La literatura estadística reconoce que los modelos de regresión para datos de conteo son adecuados cuando se busca explicar o predecir la frecuencia de eventos, como fallas, accidentes, reclamaciones, visitas o reportes, en función de variables explicativas. Cameron y Trivedi presentan una referencia clásica para el análisis de datos de conteo, señalando que estos modelos permiten representar fenómenos donde la respuesta es una frecuencia de ocurrencia [7].

La regresión de Poisson es uno de los modelos clásicos para datos de conteo y se formula dentro de la familia de los modelos lineales generalizados. Los modelos lineales generalizados permiten extender la regresión lineal tradicional a variables de respuesta con distribuciones diferentes a la normal, entre ellas la distribución Poisson para conteos [8]. En este tipo de modelo, se busca estimar el valor esperado del conteo en función de un conjunto de predictores. Sin embargo, el modelo de Poisson tiene limitaciones cuando se aplica a datos operativos reales. Una de las más relevantes es el supuesto de equidispersión, según el cual la media y la varianza del conteo son iguales. En problemas como las garantías posventa, este supuesto puede no cumplirse debido a la presencia de exceso de ceros, concentración de eventos en ciertos proyectos o categorías, heterogeneidad entre obras y variables no

observadas. En estos casos, modelos como la binomial negativa pueden ser alternativas estadísticas para manejar sobre dispersión [9].

Aunque la naturaleza de la variable objetivo justificó evaluar modelos asociados a conteos, el proyecto también consideró enfoques de aprendizaje automático más flexibles. Esto se debió a que el comportamiento de las garantías depende de interacciones complejas entre variables categóricas, estructurales, temporales e históricas. Por tanto, además de la regresión de Poisson, se evaluaron modelos basados en árboles y boosting, capaces de capturar relaciones no lineales sin exigir supuestos tan restrictivos sobre la distribución de los datos.

2.4. Modelos basados en árboles y XGBoost

Los modelos basados en árboles de decisión son ampliamente utilizados en problemas de aprendizaje supervisado con datos tabulares. Estos modelos dividen el espacio de predictores mediante reglas sucesivas, generando segmentos donde la variable objetivo puede estimarse con mayor homogeneidad. Su principal ventaja es que permiten capturar relaciones no lineales e interacciones entre variables sin requerir transformaciones paramétricas complejas.

Los métodos de ensamble fortalecen este enfoque al combinar múltiples árboles. Random Forest, por ejemplo, construye numerosos árboles sobre subconjuntos aleatorios de datos y variables, reduciendo la varianza del modelo y mejorando su estabilidad. Por otro lado, los métodos de boosting construyen árboles de manera secuencial, donde cada nuevo árbol busca corregir los errores cometidos por los anteriores. XGBoost, o Extreme Gradient Boosting, es un algoritmo de boosting basado en árboles diseñado para ser eficiente, escalable y regularizado. Chen y Guestrin lo presentan como un sistema de tree boosting escalable que incorpora optimizaciones computacionales, manejo de datos dispersos y mecanismos de regularización para mejorar el desempeño y controlar el sobreajuste [10]. Estas características explican su uso extendido en problemas prácticos de ciencia de datos con datos estructurados.

En problemas tabulares, los modelos basados en árboles continúan siendo altamente competitivos. Grinsztajn, Oyallon y Varoquaux muestran que, en conjuntos de datos tabulares típicos, los modelos basados en árboles suelen mantener un desempeño superior o comparable frente a arquitecturas de aprendizaje profundo, especialmente cuando los datos no presentan estructuras espaciales o secuenciales propias de imágenes, audio o texto [11].

Para el caso de garantías posventa, XGBoost resultó pertinente porque el fenómeno depende de variables categóricas, numéricas, temporales y de exposición. Además, permite capturar interacciones como la relación entre tipo de vivienda, cubierta, zona, macroproyecto, padre técnico e historial reciente de reclamaciones. Esta flexibilidad fue importante para representar un proceso operativo heterogéneo, donde no todas las categorías de daño ni todos los proyectos siguen el mismo comportamiento.

2.5. Calidad de datos, preparación e ingeniería de características

La calidad de los datos es un factor determinante en el desempeño de los modelos predictivos. En proyectos aplicados de ciencia de datos, el éxito del modelo depende no solo del algoritmo utilizado, sino también de la forma en que se limpian, integran, transforman y estructuran las fuentes disponibles. Errores de digitación, valores faltantes, duplicidades, cambios de nomenclatura, inconsistencias temporales o categorías no homologadas pueden afectar directamente la capacidad de aprendizaje del modelo.

En este proyecto, la preparación de datos fue especialmente relevante debido a que las fuentes provenían de sistemas operativos con cambios históricos en sus estructuras y taxonomías. Fue necesario homologar nombres de proyectos, estandarizar padres o capítulos de daño, depurar registros sin llaves de cruce, construir una fecha de entrega homologada y generar variables derivadas como meses desde entrega, exposición mensual de inmuebles e historial reciente de reclamaciones. La ingeniería de características permitió transformar las fuentes originales en variables con significado predictivo. Entre las variables derivadas más importantes se encontraron las relacionadas con exposición y comportamiento histórico. La exposición permitió representar cuántos inmuebles estaban efectivamente dentro de la ventana de garantía en cada periodo, mientras que las variables históricas permitieron capturar la persistencia temporal de las reclamaciones por proyecto y padre/capítulo.

La inclusión de observaciones con cero reclamaciones también fue una decisión clave de estructuración. Si el modelo se hubiera entrenado únicamente con combinaciones donde existieron locativas, habría aprendido solo de eventos positivos y no de escenarios donde había inmuebles expuestos sin reclamaciones. Al construir una grilla completa de exposición por padre/capítulo, el modelo pudo aprender tanto la ocurrencia como la ausencia de garantías bajo condiciones de exposición.

2.6. Metodología CRISP-DM

El desarrollo del proyecto se alineó con la metodología CRISP-DM, ampliamente utilizada en proyectos de minería de datos y ciencia de datos aplicada. CRISP-DM, sigla de Cross-Industry Standard Process for Data Mining, propone un proceso estructurado para desarrollar proyectos analíticos de manera organizada, trazable y reproducible.

Wirth y Hipp describen CRISP-DM como un modelo de proceso estándar, independiente de la industria y de la tecnología utilizada, orientado a hacer los proyectos de minería de datos más confiables, repetibles, manejables y eficientes [12]. La guía CRISP-DM 1.0 define seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación [13].

En revisiones recientes, CRISP-DM continúa siendo reconocido como un estándar de facto para proyectos de minería de datos y ciencia de datos aplicada. Schröer, Kruse y Gómez señalan que, más de dos décadas después de su publicación, el modelo sigue siendo utilizado por su

estructura clara y su capacidad para organizar proyectos analíticos en diferentes contextos [14]. En este proyecto, CRISP-DM se reflejó en las fases desarrolladas. Primero se comprendió el problema de negocio asociado a la gestión reactiva de garantías posventa. Luego se analizaron las fuentes disponibles, se prepararon los datos, se construyeron modelos predictivos, se evaluó su desempeño con métricas cuantitativas y se generaron resultados estructurados para su consulta mediante una herramienta de visualización. Esta secuencia permitió mantener trazabilidad entre el problema operativo, las decisiones metodológicas y los resultados obtenidos.

2.7. Antecedentes

El uso de modelos predictivos para garantías posventa en vivienda es un campo todavía poco explorado en comparación con otras aplicaciones de ciencia de datos en construcción, como predicción de costos, planeación de obra, seguridad laboral, mantenimiento, productividad o precios de vivienda. Sin embargo, existen antecedentes relacionados que permiten contextualizar la relevancia del problema.

En el ámbito de los defectos posentrega, Forcada et al. estudiaron defectos en viviendas después de la entrega y analizaron sus fuentes y orígenes, mostrando que los problemas posentrega pueden estar relacionados con procesos de construcción, supervisión y control de calidad [4]. Estos estudios respaldan la idea de que las reclamaciones no deben entenderse solo como casos aislados, sino como señales de procesos técnicos y operativos que pueden analizarse sistemáticamente.

Plebankiewicz, Zima y Wieczorek analizaron defectos en edificios residenciales reportados durante el periodo de garantía, lo que resulta especialmente pertinente para este proyecto, dado que se enfoca en la ocurrencia de fallas después de la entrega del inmueble [5]. Este tipo de estudios muestra que la información generada durante la garantía puede utilizarse para identificar categorías frecuentes de defectos y oportunidades de mejora en la gestión de calidad.

Un aporte a la comprensión del fenómeno de las reclamaciones en Colombia provino de un estudio académico de la Universidad Santo Tomás (2015), titulado *“Estudio y análisis de los factores que generan reclamaciones de posventa en la construcción de vivienda multifamiliar”*. Esta investigación cualitativa indagó en las causas subyacentes de las reclamaciones tras la entrega de viviendas, a través de entrevistas y revisión de casos reales. Sus hallazgos mostraron que los reclamos de los clientes no se originaban únicamente por fallas técnicas en la obra, sino también por deficiencias en la atención al cliente y en la gestión interna de las garantías por parte de las empresas. Muchos problemas podrían haberse evitado con un mayor control de calidad durante la ejecución (prevención técnica), pero también mediante una mejor comunicación con los compradores y procesos más estructurados de seguimiento posventa (prevención en servicio). El estudio de Santo Tomás propuso estrategias preventivas como la estandarización de procesos de entrega, la implementación de mecanismos de control

de calidad en obra y la capacitación del personal encargado de posventa. Estos resultados fueron muy relevantes para nuestro proyecto, ya que subrayaron la naturaleza *multidimensional* de las reclamaciones: no son solo eventos técnicos aislados, sino hechos ligados a la percepción del cliente, a la eficiencia de la comunicación y a la cultura organizacional. Por lo tanto, cualquier modelo predictivo verdaderamente útil debía tener en cuenta este contexto amplio y ser acompañado de mejoras en procesos y capacitación, para que las predicciones se traduzcan en acciones efectivas y satisfacción del cliente. [15]

A nivel internacional, vale la pena mencionar investigaciones que combinaron enfoques de construcción *esbelta* (Lean Construction) con técnicas de inteligencia artificial. Por ejemplo, Correa et al. (2024) presentaron una revisión titulada “*Lean Construction Strategies Supported by Artificial Intelligence Techniques*”, en la cual se documentó que algoritmos de ML como las redes neuronales artificiales y las máquinas de soporte vectorial mejoraron significativamente la gestión de proyectos en construcción (optimización de cronogramas, control de costos, etc.). No obstante, estos autores señalaron que persistían limitaciones comunes en el sector: la baja calidad de los datos disponibles, la escasez de estudios aplicados específicamente a construcción, y barreras culturales dentro de las empresas para adoptar soluciones basadas en datos. Estos obstáculos explicaban en parte por qué la analítica predictiva posventa no había sido ampliamente implementada todavía, subrayando la necesidad de mejorar la recolección de datos y promover un cambio de mentalidad en las organizaciones constructoras [16]

Finalmente, las revisiones recientes sobre inteligencia artificial y aprendizaje automático en construcción muestran que estas técnicas tienen un potencial creciente para apoyar la toma de decisiones durante el ciclo de vida de los proyectos [6]. Sin embargo, también advierten retos asociados a calidad de datos, integración de fuentes y adopción organizacional. Estos retos fueron consistentes con los desafíos encontrados en este proyecto, especialmente en la homologación de información histórica, la construcción de variables confiables y la necesidad de convertir resultados analíticos en herramientas comprensibles para el negocio.

3. PREPARACIÓN Y ESTRUCTURACIÓN DE LOS DATOS

Este capítulo presenta el proceso de preparación, depuración, integración y estructuración de los datos utilizados para el desarrollo del modelo predictivo de garantías posventa. La preparación de datos constituyó una etapa central del proyecto, debido a que las fuentes disponibles provenían de sistemas operativos con estructuras, nomenclaturas y niveles de completitud diferentes, por lo cual, antes del modelado fue necesario transformar las bases originales en un conjunto analítico homogéneo, trazable y compatible con técnicas de aprendizaje automático supervisado.

El proceso incluyó la limpieza de registros, la homologación de nombres de proyectos, la estandarización de categorías de daño, la integración de entregas y garantías a nivel de inmueble, la incorporación de características técnicas de los proyectos, la construcción de la exposición mensual de inmuebles y la generación de una base de modelado con observaciones positivas y ceros estructurales. Esta última decisión fue fundamental, ya que el modelo no solo debía aprender a partir de los casos donde sí se presentaron reclamaciones, sino también de los escenarios en los que existían inmuebles expuestos y no se registraron garantías. La preparación se realizó mediante un flujo reproducible en Python, con el fin de garantizar consistencia entre la base histórica usada para entrenamiento y la base futura utilizada para predicción. Esta trazabilidad permitió que las transformaciones aplicadas a los datos históricos pudieran replicarse posteriormente sobre el archivo de entregas futuras.

3.1. Fuentes de datos utilizadas

Para el desarrollo del modelo se utilizaron fuentes internas de la constructora y tablas manuales de apoyo construidas para homologar y enriquecer la información. Las fuentes principales fueron las siguientes:

- **Base de entregas históricas.** Esta fuente contiene los registros de inmuebles entregados, con información de proyecto, inmueble y fechas asociadas al proceso de entrega. Las variables centrales fueron *Proyecto*, *Inmueble*, *Fecha de creación*, *Fecha firma acta (FEPE)*, *Fecha Programación Entrega* y *Fecha de Entrega*. Esta base permitió identificar el momento a partir del cual cada inmueble entró en periodo de garantía.

Dado que no todos los registros contaban con la fecha oficial de entrega diligenciada, se definió una jerarquía de fechas para construir una fecha homologada de entrega: En primer lugar, se utilizó *Fecha de Entrega*; si esta se encontraba vacía, se tomó *Fecha Programación Entrega*; si esta tampoco estaba disponible, se usó *Fecha firma acta (FEPE)*; y, finalmente, en ausencia de las anteriores, se empleó *Fecha de creación*. En los casos donde un mismo proyecto e inmueble presentaba más de una fecha de entrega, se conservó la fecha más reciente, considerando que esta podía corresponder a correcciones de registro o a procesos de reventa del inmueble, según se consultó con los expertos en el negocio.

Esta base de datos de entregas se identifica dentro del sistema como *Entregas_historicas.xlsx* y se extrae desde el sistema Microsoft CRM de la empresa.

- **Información de nivel de acabados entregados en el inmueble:** El archivo *negocios_activos.xlsx*, actualizado el 10 de julio de 2025 a las 2:45 p. m., corresponde a la base comercial donde se registran las condiciones de venta de cada inmueble. Aunque incluye múltiples campos administrativos y financieros, el propósito de su incorporación al modelo predictivo es exclusivamente identificar el tipo de acabado con el cual fue entregado cada inmueble.

Dado que los negocios se consolidan aproximadamente 18 meses antes de la entrega, la información relevante para este proyecto —particularmente el campo *Tipo de acabado*— se considera completa, aun cuando la fecha de actualización sea anterior a algunas entregas recientes. Para enlazar esta tabla con la base de garantías, se utilizan las columnas *Proyecto + Inmueble*, las cuales funcionan como llaves compuestas que permiten integrar variables comerciales al dataset técnico-operativo.

La variable de interés *Tipo de acabado* se homologó en dos categorías principales: Con Acabados y Sin Acabados, correspondientes a T2 y ON respectivamente, mientras que categorías intermedias llamadas T0 y T1 se aproximaron al equivalente más cercano ya que su volumen era mínimo y no representaba valor tener como variables independientes.

Para proyectos del Tipo VIS y CIUDADELA, se imputó directamente Sin Acabados de acuerdo con las reglas de negocio, mientras que para proyectos NO VIS, el acabado se validó contra la base de negocios activos. En los casos donde no fue posible determinar el acabado, se conservó la categoría *Desconocido*, evitando inferir esta variable a partir de la cantidad de garantías para no introducir fuga de información.

- **Registro de garantías:** El archivo *locativas_garantías.xlsx*, consultado el 18 de octubre de 2025 a la 1:53 p. m., constituye la fuente central del proyecto, ya que contiene el registro detallado de cada garantía posventa reportada por los clientes. Sus variables permiten identificar el tipo de daño, el espacio afectado, la clasificación técnica de la falla, la fecha en que se reportó, la fecha de atención y el estado final del caso. Asimismo, incluye información estructural que permite relacionar cada reclamación con su proyecto, su inmueble y su agrupación en macroproyectos. Para el modelo, la variable *Padre* fue especialmente importante, ya que permitió representar la categoría principal de la reclamación, mientras que variables como *Elemento*, *Daño* y *Espacio* sobre especificaban la falla reduciendo la posibilidad de generar predicción.

Los registros sin inmueble fueron excluidos del modelo principal, debido a que corresponden principalmente a reportes de zonas comunes, que si bien hacen parte del modelo de garantías que se debe ofrecer a los clientes, su exclusión se definió como una delimitación metodológica, ya que dichos reportes no cuentan con una estructura de exposición comparable a la de los inmuebles privados entregados y no existía suficiente información estructurada para modelarlos de forma consistente.

La columna Padre presentó un cambio significativo de estructura en mayo de 2024, por lo que fue necesario realizar un proceso de homologación a través de la tabla de apoyo *homologacion_padres.xlsx*, con el fin de consolidar categorías históricas de manera coherente.

- **Equivalencia de proyecto – macroproyecto:** Esta tabla permitió consolidar diferentes nombres o etapas de proyectos bajo un mismo *ProyectoMacro*. Su uso fue necesario porque las bases operativas pueden registrar un mismo proyecto con variaciones de nomenclatura, etapas, sectores o denominaciones internas. Al homologar los proyectos se redujo la cardinalidad de la variable y se facilitó el análisis agregado por macroproyecto.

Adicionalmente, cada macroproyecto incorpora su clasificación VIS/No VIS/Ciudadela y su tipología CASA/APTO, atributos necesarios para interpretar diferencias constructivas relevantes, como la presencia de cubiertas en casas o en apartamentos del último piso.

- **Caracterización del proyecto:** Esta fuente aportó características estructurales y contextuales de cada macroproyecto, tales como *TipoProyecto*, *TipoVivienda*, *CantidadApartamentos*, *LATITUD*, *LONGITUD*, *TipoCubierta*, *Pisos* y *Zona*. Estas variables permitieron enriquecer el modelo con atributos físicos, constructivos y geográficos de los proyectos.

Para la variable Pisos, los proyectos de casas o viviendas unifamiliares fueron codificados como un piso, dado que cada unidad cuenta con cubierta propia. Esta decisión fue relevante porque las garantías relacionadas con cubiertas pueden comportarse de manera diferente en proyectos de casas frente a edificios, donde solo los apartamentos del último piso están directamente expuestos a la cubierta. Las coordenadas del proyecto se pretendían utilizar para asignar la estación IDEAM más cercana mediante el algoritmo BallTree, sin embargo, los resultados no fueron certeros y se descartó, siendo reemplazada por una base de datos de sector proporcionada por el negocio.

- **Equivalencias de padres:** Debido a cambios históricos en la taxonomía de daños del sistema, fue necesario crear una tabla de equivalencias para estandarizar la variable *Padre*. Esta tabla permitió consolidar categorías que representaban el mismo concepto técnico, pero que aparecían con diferentes nombres en la base histórica. Además, se aplicó una normalización robusta de texto para evitar diferencias por mayúsculas, tildes, espacios dobles o variaciones menores de escritura.

Durante la revisión del modelo se identificó que la categoría histórica FISURAS dejó de registrarse como padre independiente y pasó a quedar contenida operativamente dentro de MUROS. Por esta razón, para efectos del modelado predictivo se integró FISURAS dentro de MUROS, manteniendo consistencia con la taxonomía reciente utilizada por la operación.

- **Archivo de entregas futuras.** Se solicitó al negocio un archivo de entregas proyectadas, utilizado para generar la predicción futura. Este archivo incluye información por inmueble, con campos de *proyecto*, *inmueble*, *tipo de acabado*, *fecha estimada de entrega*, *tipo de*

proyecto y tipo de vivienda, estimado según los procesos de negociación que maneja la empresa, utilizado para la definición de sus metas financieras y operativas del año. A partir de esta fuente se generó la exposición futura de inmuebles y se aplicó el modelo seleccionado para estimar las garantías esperadas.

- **Datos climáticos históricos.** Como fuente contextual externa, se consideraron datos climáticos asociados a precipitación histórica. Esta información se exploró debido a que ciertas categorías de reclamaciones posventa, especialmente las relacionadas con cubiertas, humedades, filtraciones y goteras, pueden verse influenciadas por condiciones de lluvia o estacionalidad climática. Para este propósito se revisaron registros de precipitación por fecha de observación, municipio y coordenadas geográficas, con el fin de evaluar la posibilidad de relacionar los proyectos con condiciones climáticas cercanas a su ubicación.

No obstante, durante el desarrollo del proyecto se identificaron limitaciones de cobertura, completitud y consistencia temporal que impedían utilizar esta fuente como variable central del modelo operativo final. Por esta razón, la información climática fue tratada como una fuente exploratoria y contextual, y no como un predictor principal en la versión final del modelo, lo que permitió mantener un flujo de modelado más estable y reproducible, evitando introducir ruido asociado a registros climáticos incompletos.

- **Índice ENSO (El Niño / La Niña):** También se consideró una fuente mensual asociada al estado climático global ENSO, con información sobre anomalías y probabilidades de ocurrencia de condiciones El Niño, Neutral y La Niña. Esta fuente fue evaluada como una posible variable contextual para representar condiciones climáticas generales que podrían incidir indirectamente en el comportamiento de algunas reclamaciones sensibles al agua.

Sin embargo, al igual que la información de precipitación, el índice ENSO no fue incorporado como componente principal del modelo final debido a que su escala mensual y global ofrecía contexto climático, pero no necesariamente explicaba de manera directa la variabilidad específica de las reclamaciones, por tanto, se dejó documentado como insumo explorado y como una línea de fortalecimiento para futuras versiones del modelo, especialmente si se logra consolidar una fuente climática más robusta, local y consistente.

Cabe destacar que todas las bases de datos internas (reclamaciones, inmuebles, proyectos, macroproyectos) fueron proporcionadas por la constructora en cuestión. Se realizaron las uniones correspondientes entre estas fuentes a través de identificadores comunes (como el código de proyecto y el identificador de inmueble) para consolidar en una sola tabla toda la información relevante por cada reclamación.

Por razones de confidencialidad asociadas a la información interna de la organización, los archivos fuente utilizados para el desarrollo del proyecto no se incluyen como anexos del presente documento. No obstante, la Tabla 1 resume las principales fuentes empleadas, su contenido y propósito dentro del proceso de modelado, diferenciando las bases internas, las tablas auxiliares de homologación y las fuentes externas exploradas como contexto climático.

Tabla 1. Fuentes de datos utilizadas

Fuente de datos	Tipo de fuente	Variables principales utilizadas	Uso dentro del proyecto
Entregas históricas	Interna	Proyecto, Inmueble, Fecha de Entrega, Fecha Programación Entrega, Fecha firma acta (FEPE), Fecha de creación	Permite identificar los inmuebles entregados y construir la fecha homologada de entrega, punto de inicio para calcular la exposición y los meses desde entrega.
Locativas de garantías	Interna	Proyecto, Inmueble, Nombre, Fecha de creación, Padre, Elemento, Daño, Espacio, ¿Aplica Garantía?	Fuente principal de reclamaciones posventa. Permite identificar cada locativa, su fecha de reporte y la categoría técnica de la garantía.
Homologación de macroproyectos	Tabla auxiliar	Proyecto, ProyectoMacro, TipoProyecto, TipoVivienda	Permite consolidar etapas o variaciones nominales de proyectos bajo un mismo macroproyecto y enriquecer la segmentación del modelo.
Especificaciones de proyectos	Tabla auxiliar	ProyectoMacro, TipoProyecto, TipoVivienda, CantidadApartamentos, LATITUD, LONGITUD, TipoCubierta, Pisos, Zona	Aporta características estructurales, constructivas y geográficas de los proyectos.
Homologación de padres/capítulos	Tabla auxiliar	PADRE, PADRE EQUIVALENTE	Permite estandarizar la taxonomía de daños y consolidar categorías históricas equivalentes.
Negocios activos	Interna	Proyecto, Inmueble, Tipo de acabado, Tipo de proyecto	Permite identificar si el inmueble fue entregado con o sin acabados.
Datos climáticos históricos	Externa/contextual	FechaObservacion, ValorObservado, Municipio, Latitud, Longitud	Fuente explorada para evaluar posibles relaciones entre lluvia y reclamaciones sensibles al agua, como cubiertas, humedades y filtraciones.
Índice ENSO	Externa/contextual	Fecha, anomalía, estado, probabilidad de El Niño, Neutral y La Niña	Fuente explorada para representar condiciones climáticas generales asociadas a fenómenos El Niño, La Niña o Neutral.
Entregas futuras	Interna/proyectada	Proyecto, Inmueble, TipoAcabado, FechaEntregaEstimada, TipoProyecto, TipoVivienda	Permite construir la exposición futura y generar predicciones de garantías esperadas para periodos posteriores.

Fuente: Elaboración propia

3.2.Limpieza, depuración y homologación

El proceso de limpieza inició con la selección de columnas relevantes en cada fuente, donde se descartaron campos administrativos, identificadores técnicos del sistema, columnas duplicadas o variables que no aportaban información útil para el objetivo predictivo. Esta reducción permitió concentrar el análisis en las variables necesarias para identificar el inmueble, caracterizar el proyecto, ubicar temporalmente la entrega y clasificar técnicamente la reclamación.

En la base de entregas, los campos de *proyecto* e *inmueble* fueron tratados como llaves principales para relacionar cada unidad entregada con posibles garantías posteriores. Se excluyeron registros sin inmueble, ya que no permitían construir una relación confiable entre entrega, exposición y reclamación, para posteriormente homologar la fecha de entrega utilizando la jerarquía definida y se conservaron únicamente registros con fecha válida.

En la base de locativas de garantías se excluyeron registros *sin proyecto, sin inmueble o sin identificador único de locativa*. Las locativas marcadas como *Atención comercial* fueron tratadas como equivalentes a garantías aplicadas, dado que, aunque no necesariamente correspondían a una obligación legal estricta, sí representaban una atención efectiva concedida al cliente y, por tanto, una carga operativa real para la posventa. Además, por su bajo volumen no era pertinente que sean tratadas como una variable independiente.

La homologación de proyectos se realizó mediante el cruce con la tabla auxiliar de *macroproyectos*. Esta etapa permitió consolidar etapas o nomenclaturas diferentes bajo un mismo *ProyectoMacro*, asegurando coherencia en el análisis histórico y evitando que variaciones nominales fueran tratadas como proyectos distintos.

La homologación de padres o capítulos se realizó mediante una clave normalizada de texto. Para ello, tanto la base de locativas como la tabla auxiliar fueron transformadas a mayúsculas, sin tildes, sin espacios múltiples y sin caracteres innecesarios. Esta decisión permitió corregir problemas como categorías equivalentes separadas por diferencias mínimas de escritura. El resultado fue la variable *PadreHomologado*, utilizada como categoría técnica principal del daño dentro del modelo.

La *Tabla 2* presenta los principales criterios de depuración y exclusión aplicados durante la preparación de los datos. Estas decisiones fueron definidas para garantizar que el conjunto de modelado representara únicamente escenarios con información suficiente para calcular exposición, tiempo desde entrega y reclamaciones asociadas a inmuebles privados.

Tabla 2. Criterios de depuración y exclusión metodológica

Criterio aplicado	Fuente afectada	Tratamiento realizado	Justificación metodológica
Registros sin proyecto	Entregas y locativas	Exclusión	No es posible asociar el registro a un macroproyecto ni enriquecerlo con variables estructurales.
Registros sin inmueble	Locativas	Exclusión del modelo principal	Corresponden principalmente a zonas comunes. No cuentan con estructura de exposición comparable con inmuebles privados.
Locativas sin identificador único	Locativas	Exclusión	No permiten garantizar unicidad del registro ni evitar duplicidades.
Duplicados por identificador de locativa	Locativas	Conservación de un único registro	Evita sobrecontar una misma garantía en la variable objetivo.
Inmuebles con varias fechas de entrega	Entregas	Conservación de la fecha más reciente	Puede reflejar correcciones de registro o reventa del inmueble. Se conserva la fecha más representativa para la garantía.
Entregas sin fecha homologada válida	Entregas	Exclusión	No se puede calcular el inicio de exposición ni los meses desde entrega.
Locativas sin entrega asociada	Integración entregas-garantías	Exclusión del modelo principal	Principalmente corresponden a inmuebles entregados antes del periodo disponible en la base de entregas. No es posible calcular exposición.
Reportes antes de la entrega	Integración entregas-garantías	Exclusión	No corresponden a garantías posentrega en sentido estricto.
Reportes posteriores a la ventana modelable	Integración entregas-garantías	Exclusión del modelo principal	Pueden corresponder a reincidencias, atenciones posteriores o situaciones fuera del alcance principal de garantía posentrega.
Reportes de zonas comunes	Locativas	Exclusión metodológica	Requieren un modelo diferente, con inventario de zonas comunes y exposición propia.
Variables climáticas con cobertura insuficiente	Datos climáticos e índice ENSO	No incorporación como predictores principales	Se exploraron como contexto, pero no se integraron al modelo final por limitaciones de completitud, cobertura y consistencia temporal.
Tipo de acabado faltante en NO VIS	Negocios activos	Categoría Desconocido cuando no había evidencia suficiente	Evita imputar acabado usando la cantidad de garantías, lo cual generaría fuga de información.

Fuente: Elaboración propia

3.3. Integración de entregas, garantías y características del proyecto

Una vez depuradas las fuentes principales, se integraron las locativas de garantías con la base de entregas históricas a través de la combinación *Proyecto e Inmueble*. Este cruce permitió asociar cada reclamación con la fecha de entrega del inmueble correspondiente y calcular el tiempo transcurrido entre la entrega y el reporte de la garantía.

A partir de esta integración se clasificaron los registros en cuatro estados metodológicos: *reportes dentro de la ventana modelable*, *reportes posteriores a la ventana de garantía*, *reportes antes de la entrega* y *reportes sin entrega asociada*. Los registros sin entrega asociada correspondían principalmente a inmuebles entregados antes del periodo disponible en la base histórica de entregas, por lo que fueron excluidos del modelo principal, debido a que no era posible calcular su exposición ni los meses transcurridos desde la entrega.

Los reportes realizados antes de la entrega también fueron excluidos, ya que no es posible considerar una garantía previa a la entrega del inmueble y que determina que constituyen un error de registro u otra naturaleza que debe ser explorada por el negocio y no hace parte del análisis de este proyecto. Por su parte, los reportes posteriores a 12 meses se consideraron fuera del alcance del modelo principal, aunque estos casos pueden representar situaciones estructurales, reincidencias o atenciones posteriores, se excluyeron para evitar ruido y mantener el foco en la ventana operativa de garantías posentrega.

Posteriormente, la base modelable se enriqueció con las *especificaciones de proyecto*, incorporando variables como *tipo de proyecto*, *tipo de vivienda*, *cantidad de apartamentos*, *tipo de cubierta*, *pisos*, *zona* y *coordenadas*. Asimismo, se incorporó el *tipo de acabado por inmueble*, aplicando reglas de negocio para imputar o clasificar valores faltantes de forma controlada.

La Tabla 3 presenta las variables finales utilizadas en el dataset de modelado. Estas variables integran información estructural del proyecto, exposición de inmuebles, temporalidad posentrega, categoría técnica del daño y comportamiento histórico reciente.

Tabla 3. Variables finales utilizadas en el dataset de modelado

Variable	Tipo	Descripción	Rol en el modelo
ProyectoMacro	Categórica	Nombre homologado del macroproyecto.	Predictor. Captura diferencias estructurales y operativas entre proyectos.
PeriodoExposicion	Temporal	Mes en el que el inmueble se encuentra expuesto a presentar garantía.	Variable de agregación y control temporal.
AnioExposicion	Numérica/temporal	Año del periodo de exposición.	Predictor temporal.
MesCalendarioExposicion	Numérica/temporal	Mes calendario del periodo de exposición.	Predictor de estacionalidad mensual.
MesesDesdeEntrega	Numérica	Número de meses transcurridos desde la entrega del inmueble.	Predictor clave de la curva posentrega.
PadreHomologado	Categórica	Categoría técnica homologada del daño.	Predictor. Representa la tipología principal de garantía.
TipoProyecto	Categórica	Clasificación del proyecto, como VIS, NO VIS, CIUDADELA o COMERCIAL.	Predictor de segmentación.
TipoVivienda	Categórica	Tipo de vivienda, como APTO, CASAS o LOCALES.	Predictor constructivo.
TipoAcabadoModelo	Categórica	Clasificación final del acabado: Con Acabados, Sin Acabados o Desconocido.	Predictor técnico-comercial.
TipoCubierta	Categórica	Sistema predominante de cubierta del proyecto.	Predictor constructivo.
Zona	Categórica	Zona geográfica del proyecto.	Predictor contextual/geográfico.
CantidadApartamentos	Numérica	Número total de unidades del proyecto.	Predictor de escala del proyecto.
PisosNumerico	Numérica	Número de pisos del proyecto; casas codificadas como 1.	Predictor estructural.

EsCasa	Binaria	Identifica si el proyecto corresponde a casas.	Predictor derivado.
LATITUD	Numérica	Coordenada geográfica del proyecto.	Predictor espacial.
LONGITUD	Numérica	Coordenada geográfica del proyecto.	Predictor espacial.
InmueblesExpuestos	Numérica	Número de inmuebles entregados y expuestos en la combinación analizada.	Predictor de exposición.
LocativasProyectoMesAnterior	Numérica	Locativas del proyecto en el mes anterior.	Predictor histórico.
LocativasProyectoUltimos3Meses	Numérica	Locativas acumuladas del proyecto en los últimos tres meses previos.	Predictor histórico.
PromedioLocativasProyectoUltimos3Meses	Numérica	Promedio mensual de locativas del proyecto en los tres meses previos.	Predictor histórico.
TasaProyectoMesAnterior	Numérica	Tasa de locativas del proyecto en el mes anterior respecto a la exposición.	Predictor histórico ajustado por exposición.
LocativasPadreMesAnterior	Numérica	Locativas del mismo proyecto y padre/capítulo en el mes anterior.	Predictor histórico principal.
LocativasPadreUltimos3Meses	Numérica	Locativas acumuladas del mismo proyecto y padre/capítulo en los tres meses previos.	Predictor histórico.
PromedioLocativasPadreUltimos3Meses	Numérica	Promedio mensual del mismo proyecto y padre/capítulo en los tres meses previos.	Predictor histórico.
LocativasReales	Numérica	Cantidad real de locativas observadas para la combinación analizada.	Variable objetivo.

Fuente: Elaboración propia

3.4.Construcción de la exposición de inmuebles

Una decisión metodológica central del proyecto fue modelar las garantías en función de la exposición de inmuebles entregados. Esta lógica reconoce que, el número esperado de reclamaciones no debe interpretarse únicamente como un conteo absoluto, sino en relación con la cantidad de unidades que se encuentran efectivamente dentro de la ventana de garantía. Para construir la exposición, cada inmueble entregado fue expandido mensualmente desde el mes 0 hasta el mes 12 posterior a su entrega, de esta manera, un inmueble entregado en un periodo determinado se consideró expuesto durante los meses posteriores en los que podía generar una garantía posentrega. Esta estructura permitió calcular, para cada combinación de proyecto, periodo y meses desde entrega, el número de inmuebles expuestos.

La inclusión del mes 12 respondió a una decisión operativa del negocio; aunque la garantía formal se asocia generalmente a un periodo de 12 meses, en la práctica no se realiza un corte estrictamente al día exacto de vencimiento, como parte de las políticas de atención al cliente. Por esta razón, se consideró la ventana modelable de 0 a 12 meses inclusive, con el fin de representar de forma más realista la operación posventa.

Esta estructura de exposición fue fundamental para evitar interpretaciones sesgadas. Por ejemplo, un proyecto con 100 inmuebles entregados y 20 reportes de garantía no representa el mismo nivel de incidencia que un proyecto con 1.000 inmuebles entregados y el mismo número de reportes. Al incorporar la exposición, el modelo puede aprender patrones de reclamación considerando la base real de unidades susceptibles de generar garantías.

3.5.Definición de la variable objetivo

La variable objetivo del modelo fue **LocativasReales**, definida como la cantidad de locativas de garantía registradas para una combinación específica de periodo, proyecto, mes desde entrega y padre homologado. Esta formulación permitió transformar el problema en una tarea de regresión supervisada sobre datos agregados.

La unidad de análisis del modelo corresponde a la combinación de:

- *ProyectoMacro*
- *PeriodoExposicion*
- *MesesDesdeEntrega*
- *PadreHomologado*
- Características estructurales del proyecto
- Tipo de acabado
- Número de inmuebles expuestos

Bajo esta estructura, ***el modelo estima cuántas garantías se esperan para una tipología de daño específica, en un proyecto determinado, durante un periodo mensual y considerando el tiempo transcurrido desde la entrega.***

Además de la cantidad de locativas, se construyeron variables auxiliares como InmueblesConReclamo, TasaLocativasPorInmuebleExpuesto y TasaInmueblesConReclamo, las cuales permitieron validar la consistencia de la base y analizar la relación entre reclamaciones y exposición.

La *Tabla 4* resume la definición de la variable objetivo y la unidad de análisis del modelo. La inclusión de observaciones con cero reclamaciones fue necesaria para evitar un sesgo de entrenamiento basado únicamente en eventos positivos.

Tabla 4. Construcción de la variable objetivo y unidad de análisis

Elemento	Definición
Variable objetivo	LocativasReales
Tipo de problema	Regresión supervisada para datos de conteo
Unidad de análisis	Combinación mensual de proyecto, padre/capítulo, mes desde entrega, características del proyecto y exposición de inmuebles
Ventana modelable	Meses desde entrega 0 a 12 inclusive
Observaciones positivas	Combinaciones donde se registraron una o más locativas reales
Observaciones cero	Combinaciones donde existían inmuebles expuestos pero no se registraron locativas
Justificación de incluir ceros	Permite que el modelo aprenda tanto la ocurrencia como la ausencia de reclamaciones bajo exposición
Uso de exposición	Ajusta la predicción según la cantidad de inmuebles entregados y susceptibles de presentar garantías
Nivel de predicción	Cantidad esperada de locativas por periodo, proyecto y padre/capítulo

Fuente: Elaboración propia

3.6.Construcción del dataset final de modelado

El dataset final se construyó a partir de una grilla completa que combinó *la exposición de inmuebles con las categorías de daño modelables*. Esta decisión permitió incluir tanto combinaciones con reclamaciones observadas como combinaciones con cero reclamaciones, siendo esta ultima indispensable, ya que el modelo debía aprender no solo cuándo ocurren garantías, sino también cuándo *no se presentan reportes a pesar de existir inmuebles expuestos*.

La construcción del dataset con ceros permitió evitar un sesgo de selección que habría ocurrido si solo se entrenaba el modelo con registros positivos. En ese caso, el modelo habría aprendido únicamente de escenarios donde ya existían garantías, perdiendo capacidad para estimar adecuadamente la ausencia o baja ocurrencia de reclamaciones.

Sobre esta base se generaron variables de historial reciente, orientadas a capturar la dependencia temporal del fenómeno. Entre estas variables se incluyeron:

- locativas del padre en el mes anterior;
- locativas del padre en los últimos tres meses;
- promedio de locativas del padre en los últimos tres meses;
- locativas del proyecto en el mes anterior;
- locativas del proyecto en los últimos tres meses;
- promedio de locativas del proyecto en los últimos tres meses;
- tasa del proyecto en el mes anterior.

Estas variables permitieron incorporar memoria histórica al modelo, reconociendo que las reclamaciones posventa pueden presentar persistencia temporal, es decir, si un proyecto empieza a concentrar reportes en una categoría específica, ese comportamiento reciente puede ser una señal relevante para anticipar reclamaciones en los meses siguientes. Esta consideración es clave en un modelo que estima situaciones tan impredecibles como lo son las garantías, principalmente por el fuerte componente humano que no se tiene en cuenta: Responsables del proyecto, contratistas, proveedores, entre otros, que pueden modificar por completo el comportamiento de un proyecto a otro, a pesar de que según las características que contemplamos en este modelo sean muy similares.

El dataset final integró variables categóricas y numéricas. Las variables categóricas fueron tratadas mediante codificación en el flujo de preprocesamiento del modelo, mientras que las numéricas conservaron su interpretación directa. Esta estructura permitió entrenar modelos supervisados comparables, y posteriormente aplicar el mismo flujo sobre la base de entregas futuras.

3.7.Consideraciones metodológicas y exclusiones

Durante la preparación de los datos se definieron varias exclusiones metodológicas orientadas a preservar la consistencia del modelo.

En primer lugar, se excluyeron los reportes de zonas comunes, identificados principalmente por registros sin inmueble asociado. Esta decisión no implica que dichos reportes carezcan de importancia para la gestión posventa, sino que requieren una estructura de modelado diferente, basada en inventarios de zonas comunes, características de equipamiento y exposición distinta a la de inmuebles privados, variables sobre las cuales no se cuenta con un

registro confiable por parte de la empresa, así que podrán considerarse en una etapa posterior a este proyecto si se diseñan las estrategias adecuadas para su recolección y registro.

En segundo lugar, se excluyeron las locativas sin fecha de entrega asociada; estos casos correspondían principalmente a inmuebles entregados antes del periodo disponible en la base histórica de entregas, y al no contar con la fecha de entrega, no era posible calcular los meses desde entrega ni la exposición del inmueble, por lo que su inclusión habría introducido ruido metodológico.

En tercer lugar, se excluyeron los reportes antes de la entrega y los reportes posteriores a la ventana modelable, ya que los primeros no corresponden a garantías identificables, mientras que los segundos pueden responder a fenómenos distintos, como reincidencias, atenciones comerciales, fallas estructurales posteriores o solicitudes fuera del periodo principal de garantía.

En cuarto lugar, las variables climáticas disponibles fueron exploradas como contexto, pero no se incorporaron como eje principal del modelo operativo final. Aunque la lluvia y fenómenos como ENSO pueden influir en ciertos tipos de reclamaciones, especialmente cubiertas, humedades, filtraciones y goteras, las fuentes disponibles presentaban limitaciones de cobertura y consistencia para integrarse plenamente en esta versión del modelo. Por tanto, se dejaron como línea de trabajo futuro y no como componente central de la predicción final.

Finalmente, se evitó inferir variables explicativas a partir del resultado observado. Por ejemplo, el tipo de acabado no se imputó con base en la cantidad de garantías del inmueble, ya que esto habría generado fuga de información. En su lugar, se utilizaron reglas previas al resultado, como la clasificación por tipo de proyecto, la información de negocios activos y la categoría Desconocido cuando no existía evidencia suficiente.

En conjunto, estas decisiones permitieron construir una base de modelado consistente, reproducible y alineada con el objetivo del proyecto: Estimar el comportamiento esperado de las garantías posventa a partir de información disponible antes o durante la exposición del inmueble, evitando el uso de información posterior que pudiera sesgar el aprendizaje del modelo.

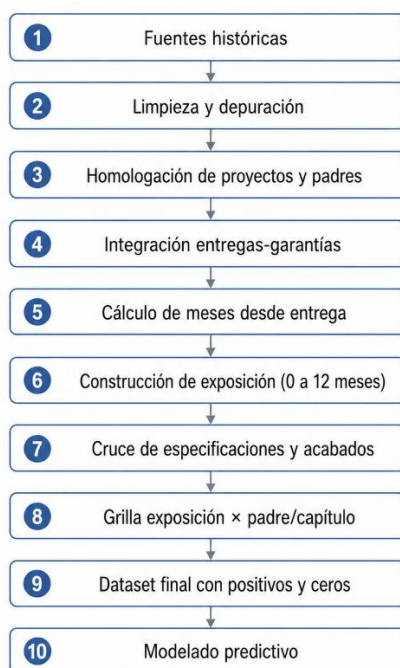
La *Tabla 5* resume las fuentes climáticas exploradas y la decisión metodológica adoptada. Aunque estas fuentes fueron consideradas desde el planteamiento inicial del proyecto, no fueron incorporadas como predictores principales en la versión final del modelo debido a limitaciones de cobertura, completitud y trazabilidad temporal.

Tabla 5. Variables climáticas exploradas y decisión metodológica

Fuente climática	Variables disponibles	Relación esperada con garantías	Decisión final
Precipitación histórica	Fecha de observación, valor observado, municipio, latitud, longitud	Posible relación con cubiertas, humedades, filtraciones y goteras.	Se exploró como fuente contextual, pero no se incorporó como predictor principal por limitaciones de cobertura y consistencia temporal.
Índice ENSO	Fecha, anomalía, estado, probabilidades de El Niño, Neutral y La Niña	Posible relación indirecta con periodos secos o lluviosos y reclamaciones sensibles al agua.	Se documentó como insumo contextual y se dejó como línea de trabajo futuro.
Coordenadas de proyectos	Latitud y longitud del macroproyecto	Permitirían asociar proyectos con fuentes climáticas cercanas.	Se conservaron como variables espaciales del modelo, pero no se integró una medición climática robusta en la versión final.

Fuente. Elaboración propia

Ilustración 1. Flujo general de preparación y estructuración de datos



Fuente. Elaboración propia

3.8.Caracterización del dataset final de modelado

Como resultado del proceso de preparación de datos descrito en las secciones anteriores, se obtuvo un dataset consolidado para modelado compuesto por 489.346 observaciones, correspondientes a combinaciones válidas de las variables de análisis definidas para el estudio. Del total de observaciones generadas, 25.156 presentaron una o más locativas registradas, mientras que 464.190 correspondieron a observaciones con valor cero, representando el 94,86 % del conjunto de datos: En total, el dataset consolidado contiene 55.774 locativas reales distribuidas entre 99 proyectos y 22 padres o capítulos homologados.

Para el entrenamiento de los modelos predictivos se utilizó el subconjunto histórico comprendido entre diciembre de 2017 y septiembre de 2025, obteniendo 450.736 observaciones efectivas de entrenamiento y un total de 55.503 locativas históricas utilizadas para el aprendizaje de los modelos.

La Tabla 6 presenta el resumen general de las principales características del dataset final de modelado.

Tabla 6. Características del dataset final de modelado

Característica	Valor
Observaciones totales del dataset consolidado	489.346
Observaciones con locativas registradas	25.156
Observaciones con valor cero	464.190
Porcentaje de observaciones con valor cero	94,86 %
Locativas reales acumuladas	55.774
Observaciones utilizadas para entrenamiento	450.736
Locativas históricas utilizadas para entrenamiento	55.503
Proyectos incluidos	99
Padres o capítulos homologados	22
Periodo cubierto por el dataset	Dic-2017 a Oct-2026
Periodo utilizado para entrenamiento	Dic-2017 a Sep-2025

Fuente: Elaboración propia.

4. MODELADO PREDICTIVO SUPERVISADO

Este capítulo presenta el proceso de modelado predictivo desarrollado para estimar la cantidad esperada de garantías posventa en proyectos residenciales de la constructora. A partir del dataset estructurado en el capítulo anterior, el problema fue formulado como una tarea de aprendizaje automático supervisado de regresión para datos de conteo, donde la variable objetivo corresponde al número de locativas reales observadas en una combinación determinada de periodo, macroproyecto, mes desde entrega y padre o capítulo técnico de la reclamación.

El enfoque de modelado buscó representar el comportamiento posventa desde una perspectiva operativa, por esta razón, las garantías no fueron tratadas únicamente como eventos individuales, sino como conteos agregados asociados a inmuebles expuestos, características del proyecto, categoría técnica del daño, tiempo transcurrido desde la entrega e historial reciente de reclamaciones. Esta formulación permitió que el modelo aprendiera tanto de los escenarios donde se presentaron garantías como de aquellos donde existían inmuebles expuestos pero no se registraron reclamaciones.

El capítulo se organiza en seis secciones: primero se presenta el enfoque general de modelado; posteriormente se describen los modelos evaluados, la estrategia de entrenamiento y validación, las métricas utilizadas, la selección del modelo final y, finalmente, la interpretación de variables del modelo seleccionado.

4.1. Enfoque general del modelado

El comportamiento de las garantías posventa presenta una naturaleza multivariada, temporal y heterogénea. La cantidad de reclamaciones puede variar según el macroproyecto, el tipo de vivienda, el tipo de proyecto, el sistema de cubierta, el nivel de acabados, la zona geográfica, la cantidad de inmuebles entregados, el tiempo transcurrido desde la entrega y el historial reciente de reportes. Por esta razón, el problema fue abordado como una tarea de regresión supervisada sobre datos tabulares.

En este enfoque, cada observación representa una combinación estructurada de variables explicativas y una variable objetivo conocida, a partir de la cual el algoritmo aprende patrones históricos que pueden utilizarse para estimar valores en escenarios futuros. La variable objetivo del modelo fue *LocativasReales*, definida como la cantidad de locativas de garantía observadas para una combinación específica de:

- macroproyecto;
- periodo mensual de exposición;
- mes transcurrido desde la entrega;

- padre o capítulo homologado;
- características estructurales del proyecto;
- tipo de acabado;
- número de inmuebles expuestos;
- comportamiento histórico reciente.

Esta formulación permitió orientar el modelo hacia una necesidad práctica del negocio: estimar cargas esperadas de garantías por periodo, proyecto y tipología de reclamación, en lugar de predecir únicamente si ocurrirá o no una garantía individual. El modelo estima el volumen esperado de locativas bajo condiciones específicas, lo cual resulta más útil para la planeación de recursos, la priorización operativa y el seguimiento preventivo de proyectos.

El uso de modelos supervisados tabulares fue pertinente por tres razones principales. En primer lugar, el fenómeno analizado combina variables categóricas, numéricas, temporales y de exposición. En segundo lugar, las relaciones entre estas variables no necesariamente son lineales; por ejemplo, el efecto del tipo de cubierta puede depender del tipo de vivienda, del número de pisos o del padre técnico analizado. En tercer lugar, el modelo debía ser compatible con una herramienta de visualización y consulta de resultados, de modo que sus predicciones pudieran convertirse en insumos operativos para la gestión posventa.

4.2. Modelos evaluados

Durante el proceso de modelado se evaluaron diferentes enfoques con el propósito de comparar su desempeño y seleccionar el modelo más adecuado para el problema. La comparación incluyó una línea base y modelos supervisados de mayor complejidad: El primer enfoque correspondió a un *baseline de tasa histórica*, construido para establecer un punto de referencia inicial. Este modelo permitió estimar las locativas esperadas a partir del comportamiento histórico promedio, sin incorporar relaciones complejas entre variables, y su función principal fue servir como criterio mínimo de comparación: cualquier modelo más avanzado debía superar este desempeño para justificar su uso.

Posteriormente se evaluó *Random Forest*, un modelo de ensamble basado en múltiples árboles de decisión; este algoritmo permite capturar relaciones no lineales y reducir la varianza mediante la combinación de árboles entrenados sobre subconjuntos de datos y variables. Su evaluación fue útil para contrastar el desempeño de un método robusto de árboles frente a modelos de boosting.

También se evaluó una *regresión de Poisson*, dado que la variable objetivo corresponde a un conteo de locativas. este tipo de modelo es conceptualmente adecuado para datos de conteo; sin embargo, su estructura puede resultar limitada cuando existen relaciones no lineales, interacciones complejas entre variables o alta heterogeneidad entre proyectos y categorías de daño.

Adicionalmente se evaluó *XGBoost Poisson*, un modelo de *gradient boosting* configurado con una función objetivo orientada a conteos. *XGBoost* construye árboles de manera secuencial, donde cada nuevo árbol busca corregir los errores residuales de los anteriores. Este enfoque permite capturar relaciones no lineales e interacciones entre variables, manteniendo mecanismos de regularización que ayudan a controlar el sobreajuste [10].

Finalmente, se desarrolló una versión de *XGBoost con variables de historial reciente*, que incorporó información rezagada del comportamiento de las locativas por proyecto y por padre/capítulo. Este modelo incluyó variables como las locativas del padre en el mes anterior, las locativas acumuladas del padre en los últimos tres meses, el promedio de locativas recientes por padre, y variables análogas a nivel de proyecto. La incorporación de estas variables permitió representar la persistencia temporal del fenómeno posventa: cuando una categoría de reclamación empieza a presentarse en un proyecto, Y su comportamiento reciente puede ser una señal relevante para estimar nuevas reclamaciones en los meses siguientes.

Durante el proceso experimental también se revisaron versiones calibradas de algunos modelos, con el objetivo de ajustar el volumen total predicho frente a periodos recientes de validación, sin embargo, la selección final se basó en el desempeño predictivo general del modelo sin calibración, priorizando el equilibrio entre error, capacidad explicativa y consistencia técnica.

4.3.Estrategia de entrenamiento y validación

La estrategia de entrenamiento y validación fue definida considerando la naturaleza temporal del problema, dado que el modelo será utilizado para estimar comportamientos futuros, por lo que la evaluación no debía realizarse mediante una división aleatoria tradicional ya que esta podría mezclar periodos antiguos y recientes tanto en entrenamiento como en prueba, generando una evaluación optimista.

Por esta razón, se adoptó una división temporal: El conjunto de entrenamiento incluyó los registros históricos observados hasta diciembre de 2024, mientras que el conjunto de prueba incluyó los meses completos observados entre enero y septiembre de 2025. El mes de octubre de 2025 fue excluido de la validación al corresponder al periodo máximo observado parcialmente en la fuente de locativas, lo cual podía introducir sesgos por incompletitud temporal. Esta estrategia permitió evaluar el desempeño del modelo en un escenario más cercano al uso real: entrenar con información pasada y validar sobre periodos posteriores no utilizados durante el entrenamiento. De esta forma, la evaluación representa mejor la capacidad del modelo para anticipar comportamientos futuros bajo condiciones operativas reales.

El conjunto histórico observado utilizado para el entrenamiento final del modelo seleccionado incluyó información hasta septiembre de 2025, y una vez seleccionado el modelo con base en la validación temporal, se entrenó una versión final utilizando todo el histórico completo

disponible, con el fin de aprovechar la mayor cantidad de información antes de generar las predicciones futuras.

4.4. Evaluación cuantitativa de desempeño

La evaluación de los modelos se realizó mediante métricas de regresión y métricas agregadas de control operativo [17]. Las métricas principales fueron:

- **MAE** o error absoluto medio.
- **RMSE** o raíz del error cuadrático medio.
- **R^2** o coeficiente de determinación.
- **MAPE sobre observaciones positivas.**
- **Total real vs. total predicho en el conjunto de prueba.**

El *MAE* mide el promedio de las diferencias absolutas entre los valores reales y los valores predichos, es una métrica de interpretación directa, ya que se expresa en las mismas unidades de la variable objetivo. En este caso, indica cuántas locativas, en promedio, se desvía el modelo para cada combinación evaluada.

El *RMSE* penaliza con mayor fuerza los errores grandes, debido a que eleva al cuadrado las diferencias antes de promediarlas. Esta métrica fue útil para identificar modelos que podían presentar errores elevados en combinaciones particulares de proyecto, padre o periodo.

El R^2 mide la proporción de variabilidad explicada por el modelo respecto a una línea base. En problemas operativos con muchos ceros, alta dispersión y variables no observadas, este indicador debe interpretarse con cautela, y por ello fue analizado junto con otras métricas.

El *MAPE sobre observaciones positivas* permitió evaluar el error porcentual únicamente en registros donde existía al menos una locativa real. Esta restricción fue necesaria porque el MAPE no es interpretable cuando el valor real es cero, y dado que el dataset contiene una gran cantidad de combinaciones sin reclamaciones, esta métrica ofreció una lectura complementaria del desempeño en escenarios donde sí hubo eventos.

Finalmente, se comparó el *total real* de locativas del periodo de prueba frente al *total predicho* por cada modelo. Esta comparación fue importante desde la perspectiva del negocio, porque la utilidad operativa del modelo depende no solo del error fila a fila, sino también de su capacidad para aproximarse a la carga total esperada de garantías. En esta evaluación final, si bien se identificaron algunas diferencias representativas, la consulta con expertos del negocio permitió confirmar que los proyectos atípicos presentaban características diferenciadoras que explican su comportamiento. Dichas características no hacen parte de las variables consideradas en el modelo, tanto en escenarios de sobrepredicción como de subpredicción.

Por lo anterior, los resultados se consideran consistentes y se reafirma la confianza en el desempeño del modelo.

La Tabla 7 presenta la comparación de los modelos evaluados durante la fase experimental. Se incluyeron modelos base, modelos de ensamble, una regresión de Poisson y versiones con variables históricas recientes. La comparación se realizó sobre el conjunto de prueba temporal, correspondiente al periodo enero-septiembre de 2025.

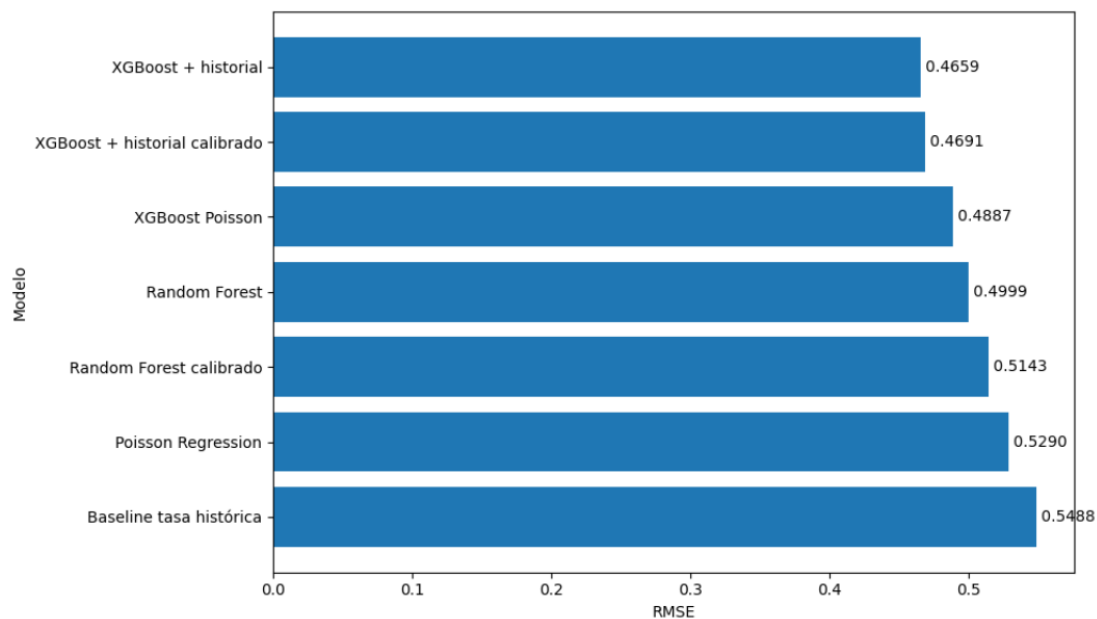
Tabla 7. Comparación de modelos evaluados

Modelo	MAE	RMSE	R ²	MAPE real positivo (%)	Total real test	Total predicho test
Baseline tasa histórica	0,127556	0,548774	-0,060550	77,598653	3.328	4.574,265082
Random Forest	0,122420	0,499931	0,119835	72,344755	3.328	4.012,588189
Poisson Regression	0,186341	0,528977	0,014588	84,236447	3.328	5.706,165593
Random Forest calibrado	0,119385	0,514309	0,068479	78,450087	3.328	3.702,137891
XGBoost Poisson	0,122627	0,488708	0,158909	71,320160	3.328	4.055,613281
XGBoost + historial	0,111872	0,465917	0,235528	68,834192	3.328	3.829,739258
XGBoost + historial calibrado	0,109652	0,469057	0,225190	72,247147	3.328	3.610,941933

Fuente: Extraído del modelo en VisualStudio

El modelo XGBoost + historial presentó el mejor equilibrio general entre las métricas evaluadas. Aunque la versión calibrada obtuvo un MAE ligeramente menor y un total predicho más cercano al total real, el modelo XGBoost + historial sin calibración alcanzó el menor RMSE, el mayor R² y el menor MAPE sobre registros positivos. Por esta razón, fue seleccionado como modelo principal para la generación de predicciones futuras.

Ilustración 2. Comparación de modelos evaluados



Nota: La ilustración 2 presenta la comparación de modelos según RMSE. Esta métrica permite identificar qué modelo presentó menor penalización por errores grandes en el conjunto de prueba temporal.

Fuente: Elaboración propia en VisualStudio

4.5. Selección del modelo final

A partir de la comparación de modelos, se seleccionó como modelo final el *XGBoost con variables de historial reciente*. Este modelo presentó el mejor equilibrio técnico general frente a las alternativas evaluadas.

En el conjunto de prueba, el modelo XGBoost + historial obtuvo los siguientes resultados:

- **MAE:** 0,111872
- **RMSE:** 0,465917
- **R²:** 0,235528
- **MAPE en registros positivos:** 68,834192 %
- **Total real en prueba:** 3.328 locativas
- **Total predicho en prueba:** 3.829,739258 locativas

Aunque la versión calibrada del modelo logró una predicción total más cercana al volumen real, presentó un desempeño inferior en algunas métricas de evaluación genera, por esta razón, se decidió conservar como modelo principal la versión no calibrada, debido a que

mostró mejor capacidad explicativa y mejor equilibrio entre error absoluto, error cuadrático y desempeño porcentual en registros con reclamaciones reales.

La selección del modelo *XGBoost + historial* también se justificó desde la perspectiva del negocio. Las garantías posventa no dependen únicamente de las características estáticas del proyecto, sino también del comportamiento reciente de las reclamaciones: La presencia de locativas en un padre/capítulo durante el mes anterior o durante los últimos tres meses puede indicar condiciones técnicas, operativas o de registro que persisten temporalmente, por lo tanto, incorporar variables de historial permitió que el modelo capturara señales relevantes del comportamiento reciente y no solo patrones promedio históricos.

El modelo final fue posteriormente entrenado con todo el histórico observado disponible hasta septiembre de 2025, con el objetivo de generar predicciones futuras a partir del archivo de entregas proyectadas. Esta versión final fue utilizada para estimar las locativas esperadas durante 2026 y los meses posteriores asociados a la ventana de exposición de las entregas futuras.

La Tabla 8 resume las métricas del modelo seleccionado. Estas métricas permiten evaluar el desempeño desde diferentes perspectivas: error promedio, penalización de errores grandes, capacidad explicativa, error porcentual en observaciones positivas y aproximación al volumen total de locativas del periodo de prueba.

Tabla 8. Métricas del modelo seleccionado

Métrica	Resultado	Interpretación
MAE	0,111872	Error absoluto promedio por combinación evaluada. Indica una desviación baja en la unidad de análisis del modelo.
RMSE	0,465917	Penaliza errores grandes. Fue el menor RMSE entre los modelos evaluados, lo que indica mejor control de errores altos.
R ²	0,235528	Mayor capacidad explicativa relativa entre los modelos evaluados. Debe interpretarse considerando la alta presencia de ceros y la variabilidad del fenómeno.
MAPE real positivo	68,834192 %	Error porcentual calculado únicamente sobre observaciones con locativas reales positivas. Fue el menor entre los modelos evaluados.
Total real test	3.328	Total de locativas observadas en el conjunto de prueba.
Total predicho test	3.829,739258	Total de locativas estimadas por el modelo en el conjunto de prueba.

Fuente: Elaboración propia

El desempeño del modelo seleccionado debe interpretarse considerando que la base de modelado contiene una alta proporción de combinaciones con cero reclamaciones y que las garantías posventa dependen de factores no completamente observables en las fuentes disponibles. Aun así, el modelo logró mejorar el desempeño de la línea base y de los modelos alternativos, especialmente al incorporar variables de historial reciente por proyecto y padre/capítulo.

A continuación, presentamos en la Tabla 9 los hiperparámetros utilizados en el modelo XGBoost seleccionado. Estos valores fueron definidos para lograr un equilibrio entre capacidad predictiva, control del sobreajuste y estabilidad del modelo sobre datos tabulares con alta presencia de ceros y comportamiento heterogéneo por proyecto y padre/capítulo.

Tabla 9. Hiperparámetros del modelo XGBoost seleccionado

Hiperparámetro	Valor utilizado	Descripción
objective	count:poisson	Función objetivo orientada a datos de conteo.
n_estimators	300	Número de árboles construidos por el modelo.
max_depth	6	Profundidad máxima de cada árbol.
learning_rate	0,05	Tasa de aprendizaje del modelo.
subsample	0,8	Proporción de observaciones usadas en cada iteración.
colsample_bytree	0,8	Proporción de variables usadas por árbol.
min_child_weight	5	Control de complejidad para evitar particiones poco robustas.
reg_alpha	0,1	Regularización L1.
reg_lambda	1,0	Regularización L2.
random_state	42	Semilla para reproducibilidad.

Fuente: Elaboración propia

4.6. Interpretación de variables del modelo

La interpretación del modelo se realizó mediante el análisis de importancia de variables. Este análisis permitió identificar qué predictores aportaron más a la estimación de locativas esperadas. Debido a que las variables categóricas fueron codificadas internamente en múltiples columnas, se revisaron dos niveles de importancia: importancia individual de variables transformadas e importancia agrupada por variable original.

A nivel individual, la variable con mayor importancia fue *LocativasPadreMesAnterior*, con una importancia de 0,251459. Esto indica que el comportamiento del mismo padre o capítulo en el mes inmediatamente anterior fue la señal predictiva más fuerte del modelo, lo que en términos de negocio sugiere que cuando un proyecto registra reclamaciones recientes en una tipología específica, existe una mayor probabilidad de que esa dinámica continúe en los meses siguientes.

Las siguientes variables más relevantes fueron *LocativasPadreUltimos3Meses*, *PromedioLocativasPadreUltimos3Meses*, *InmueblesExpuestos* y algunas categorías específicas de *PadreHomologado*, como *CUBIERTAS*, *ELECTRICO*, *VENTANAS VIDRIOS*, *MUROS* e *HIDROSANITARIO*. Este patrón confirma que el modelo combina información de comportamiento reciente, exposición y categoría técnica del daño para estimar las reclamaciones esperadas.

La Tabla 10 presenta la importancia agrupada de variables del modelo seleccionado. Esta agrupación fue necesaria porque las variables categóricas fueron transformadas internamente mediante codificación, generando múltiples columnas derivadas por cada variable original.

Tabla 10. Importancia agrupada de variables del modelo seleccionado

Variable original	Importancia total	Importancia relativa (%)
LocativasPadreMesAnterior	0,251459	25,145924
ProyectoMacro	0,203219	20,321898
PadreHomologado	0,183187	18,318726
LocativasPadreUltimos3Meses	0,083185	8,318535
PromedioLocativasPadreUltimos3Meses	0,049063	4,906310
TipoProyecto	0,039697	3,969734
Zona	0,035347	3,534730
InmueblesExpuestos	0,029570	2,956990
TipoCubierta	0,027484	2,748378
TipoAcabadoModelo	0,021679	2,167857
MesesDesdeEntrega	0,012125	1,212512
TipoVivienda	0,011701	1,170074
CantidadApartamentos	0,008357	0,835696
LocativasProyectoUltimos3Meses	0,007808	0,780769
PromedioLocativasProyectoUltimos3Meses	0,007155	0,715456
TasaProyectoMesAnterior	0,005525	0,552512
LATITUD	0,004555	0,455548
MesCalendarioExposicion	0,003595	0,359497
AnioExposicion	0,003582	0,358184
LONGITUD	0,003504	0,350367
LocativasProyectoMesAnterior	0,003406	0,340585
PisosNumerico	0,003313	0,331261
EsCasa	0,001485	0,148459

Fuente: Elaboración propia

A nivel agrupado, las variables más importantes fueron:

- LocativasPadreMesAnterior: 25,15 %
- ProyectoMacro: 20,32 %
- PadreHomologado: 18,32 %
- LocativasPadreUltimos3Meses: 8,32 %
- PromedioLocativasPadreUltimos3Meses: 4,91 %
- TipoProyecto: 3,97 %
- Zona: 3,53 %
- InmueblesExpuestos: 2,96 %
- TipoCubierta: 2,75 %
- TipoAcabadoModelo: 2,17 %

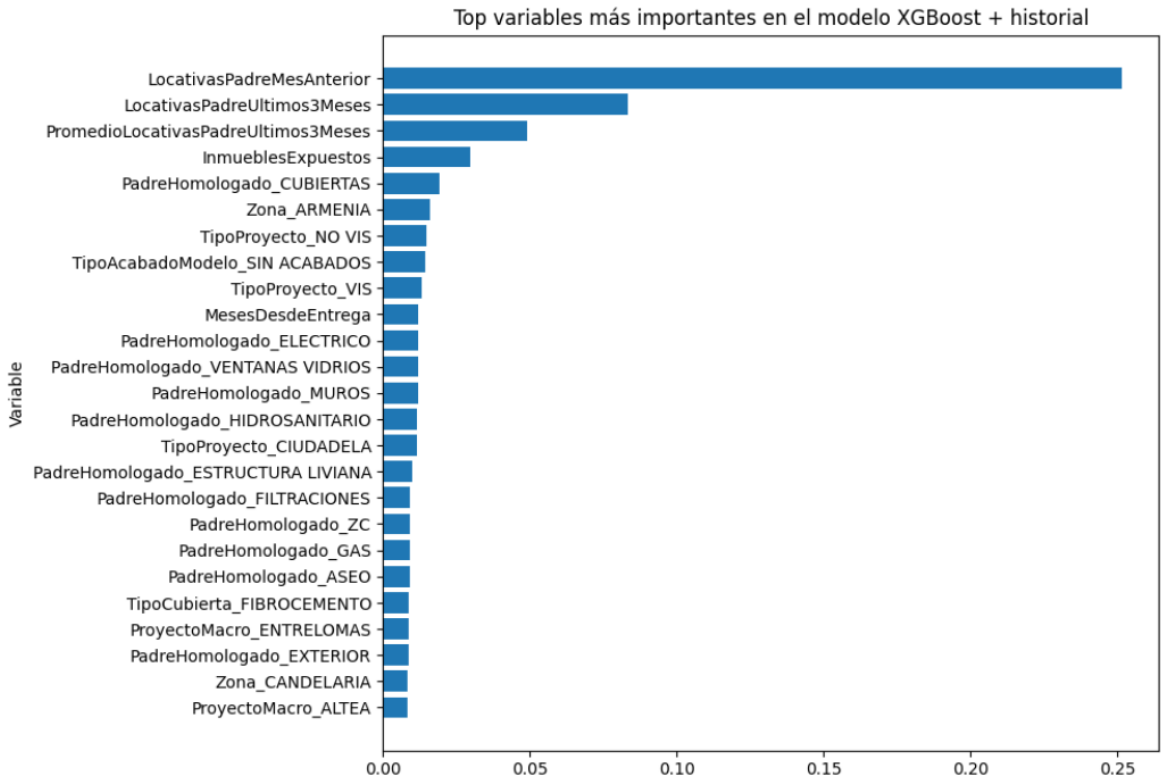
Estos resultados evidencian que el modelo no predice únicamente a partir de características estáticas del proyecto, sino que el mayor peso lo tienen las variables de comportamiento histórico reciente por padre o capítulo, seguidas por el macroproyecto y la categoría técnica de la reclamación.

La importancia de *ProyectoMacro* indica que existen diferencias relevantes entre proyectos que no son explicadas completamente por las demás variables disponibles, estas diferencias pueden estar asociadas a factores no observados en la base, como contratistas, prácticas constructivas, etapas de obra, condiciones específicas de entrega, calidad de materiales o dinámicas particulares de atención posventa.

La importancia de *PadreHomologado* confirma que las categorías técnicas de daño tienen comportamientos diferenciados. No todas las reclamaciones se presentan con la misma frecuencia ni siguen la misma curva temporal, por ejemplo, categorías como cubiertas, muros, hidrosanitario, ventanas y eléctrico pueden responder a dinámicas técnicas distintas y requerir estrategias de seguimiento diferenciadas.

Las variables estructurales, como *TipoProyecto*, *Zona*, *TipoCubierta*, *TipoAcabadoModelo*, *TipoVivienda*, *CantidadApartamentos* y *PisosNumerico*, también aportaron al modelo, aunque con menor peso relativo frente al historial reciente. Esto no significa que carezcan de importancia, sino que, una vez el modelo conoce el proyecto, el padre/capítulo y el comportamiento reciente, estas variables ayudan a ajustar la predicción.

Ilustración 3. Importancia agrupada de variables



Fuente: Elaboración propia en VisualStudio

Es importante señalar que la importancia de variables no debe interpretarse como causalidad directa: Una variable con alta importancia indica que aportó información útil para mejorar la predicción, pero no necesariamente que sea la causa de las reclamaciones, por tanto, los resultados del modelo deben entenderse como señales predictivas que apoyan la toma de decisiones, y deben complementarse con el conocimiento técnico del negocio antes de definir acciones correctivas o preventivas.

En síntesis, el modelo seleccionado combina tres dimensiones principales: exposición de inmuebles, características del proyecto y comportamiento histórico reciente. Esta integración permitió representar de manera más realista el fenómeno posventa y generar estimaciones útiles para la planeación operativa.

5. RESULTADOS Y PREDICCIÓN FUTURA

Este capítulo presenta los resultados obtenidos al aplicar el modelo predictivo seleccionado sobre las entregas futuras proyectadas. Mientras el capítulo anterior se concentró en la construcción, evaluación y selección del modelo, este capítulo se enfoca en la interpretación aplicada de las predicciones generadas y en su utilidad para la gestión posventa.

El modelo seleccionado, XGBoost con variables de historial reciente, fue utilizado para estimar la cantidad esperada de locativas de garantía a partir de la exposición futura de inmuebles. Para ello, se tomó el archivo de entregas futuras, se expandió cada inmueble durante la ventana de exposición de 0 a 12 meses posteriores a su entrega estimada, y se generaron combinaciones por periodo, macroproyecto y padre/capítulo homologado, y finalmente el modelo estimó la cantidad esperada de locativas para cada combinación.

Los resultados se presentan en cuatro niveles de análisis: *predicción total para 2026*, *distribución mensual*, *concentración por proyecto* y *concentración por padre o capítulo técnico*. Esta estructura permite interpretar la carga futura esperada desde una perspectiva operativa, facilitando la identificación de periodos, proyectos y tipologías de reclamación que podrían requerir mayor seguimiento.

5.1. Justificación del horizonte de aplicación futura

Las predicciones presentadas en esta sección corresponden a un escenario de aplicación futura del modelo y no a un proceso adicional de validación, ya que la evaluación del desempeño fue realizada exclusivamente sobre información histórica disponible, mediante la separación de conjuntos de entrenamiento y prueba descrita en el capítulo anterior.

El horizonte correspondiente al año 2026 fue seleccionado debido a que representaba el periodo futuro más próximo para el cual la organización disponía de información estructurada sobre entregas proyectadas y características asociadas a los proyectos. Esta información permitió construir la exposición futura de inmuebles y generar las variables requeridas por el modelo para realizar estimaciones de carga posventa.

En consecuencia, las predicciones obtenidas para 2026 deben interpretarse como una aplicación prospectiva del modelo bajo las condiciones observadas históricamente y las características conocidas de las entregas futuras. Dado que al momento del desarrollo del proyecto no existían observaciones reales correspondientes a dicho periodo, estas estimaciones no pueden ser validadas directamente y su propósito principal consiste en apoyar procesos de planeación operativa, asignación de recursos y priorización de actividades dentro del área de posventa.

5.2. Aplicación del modelo sobre entregas futuras

Para generar las predicciones futuras se utilizó el modelo *XGBoost con historial reciente* entrenado con todo el histórico observado disponible. La base futura se construyó a partir de 4.306 entregas proyectadas, las cuales fueron expandidas durante los meses posteriores a la entrega estimada para representar la exposición de los inmuebles dentro de la ventana de garantía. La exposición futura se construyó bajo la misma lógica utilizada en el entrenamiento histórico, cada inmueble fue considerado expuesto desde el mes 0 hasta el mes 12 posterior a la entrega, y a continuación esta exposición fue combinada con los padres o capítulos homologados modelables, con el fin de estimar la cantidad esperada de locativas por tipo de reclamación.

Dado que el modelo seleccionado incorpora variables de historial reciente, la predicción futura se realizó de forma recursiva mes a mes, y esto significa que, para cada periodo futuro, el modelo utilizó el historial real disponible hasta el último mes observado y a medida que avanzaba en los meses futuros incorporó sus propias predicciones como referencia para calcular las variables rezagadas de los periodos siguientes. Esta estrategia permitió conservar la lógica temporal del modelo, evitando asumir que los meses futuros no tienen relación entre sí.

El resultado general de este proceso fue una estimación total de aproximadamente 4.247 locativas esperadas para el año 2026. Este valor representa una estimación agregada de la carga futura de garantías, considerando las entregas proyectadas, la exposición mensual de inmuebles, las características de los proyectos y el comportamiento histórico reciente de las reclamaciones.

5.3. Predicción mensual de locativas para 2026

La predicción mensual permite observar cómo se distribuye la carga esperada de garantías a lo largo del año 2026. Este análisis es relevante para la *planeación operativa*, dado que la demanda de atención posventa no necesariamente se distribuye de manera uniforme durante todos los meses: Los resultados muestran que la cantidad esperada de locativas se mantiene entre aproximadamente 287 y 495 casos mensuales, siendo el menor valor estimado en marzo de 2026, con cerca de 287 locativas, mientras que el mayor valor se presenta en noviembre de 2026, con aproximadamente 495 locativas. Esta variación está relacionada con la acumulación progresiva de inmuebles expuestos y con el comportamiento esperado de las reclamaciones según el tiempo transcurrido desde la entrega.

La Tabla 11 presenta la predicción mensual de locativas para 2026. Los resultados muestran una carga esperada creciente hacia el último trimestre del año, con el valor máximo en noviembre de 2026.

Tabla 11. Predicción mensual de locativas para 2026

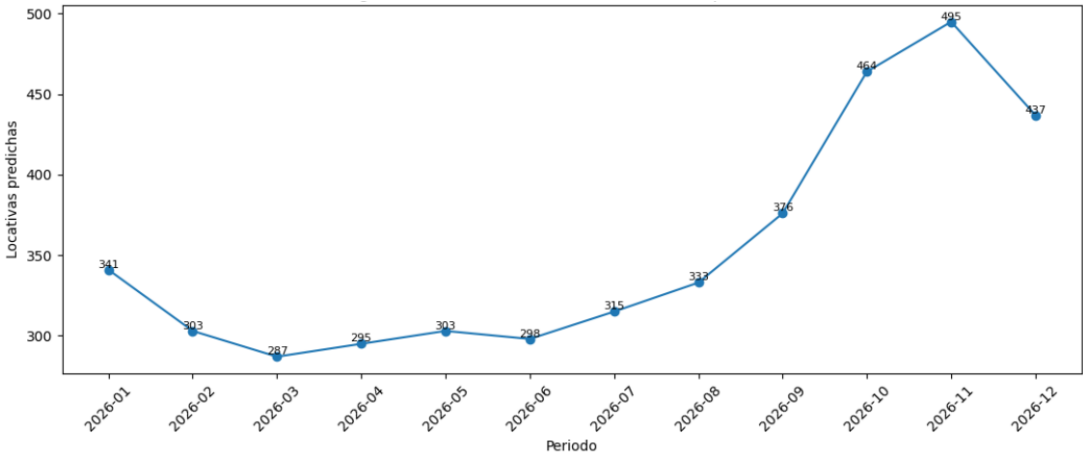
Periodo	Locativas predichas	Inmuebles expuestos	Locativas redondeadas
2026-01	341,256104	1.537	341
2026-02	302,660645	1.615	303
2026-03	286,958435	1.788	287
2026-04	294,931488	1.896	295
2026-05	302,635712	2.052	303
2026-06	298,472626	2.142	298
2026-07	315,453979	2.231	315
2026-08	333,432648	2.438	333
2026-09	376,314911	2.606	376
2026-10	463,768402	3.207	464
2026-11	494,605225	3.336	495
2026-12	436,948761	3.480	437
Total 2026	4.247,439000	—	4.247

Fuente: Elaboración propia

Es importante aclarar que la variable *InmueblesExpuestos* no representa únicamente los inmuebles entregados durante cada mes, sino la cantidad de unidades que se encuentran dentro de la ventana de exposición en ese periodo. Por esta razón, el número de inmuebles expuestos tiende a aumentar a medida que se acumulan entregas recientes que continúan dentro del periodo de garantía.

Desde el punto de vista operativo, los meses de octubre, noviembre y diciembre de 2026 requieren especial atención, ya que concentran las mayores estimaciones de locativas: Este resultado puede utilizarse como insumo para anticipar necesidades de personal, programación de visitas técnicas, coordinación con contratistas y seguimiento preventivo a categorías críticas.

Ilustración 4. Predicción mensual de locativas para 2026



Fuente: Elaboración propia en VisualStudio

5.4.Resultados por proyecto

El análisis por macroproyecto permite identificar dónde se concentra la mayor carga futura esperada de garantías, siendo una vista particularmente útil para la gestión posventa, ya que permite priorizar proyectos que podrían requerir mayor capacidad de atención o seguimiento preventivo.

Tabla 12.Proyectos con mayor predicción de locativas en 2026

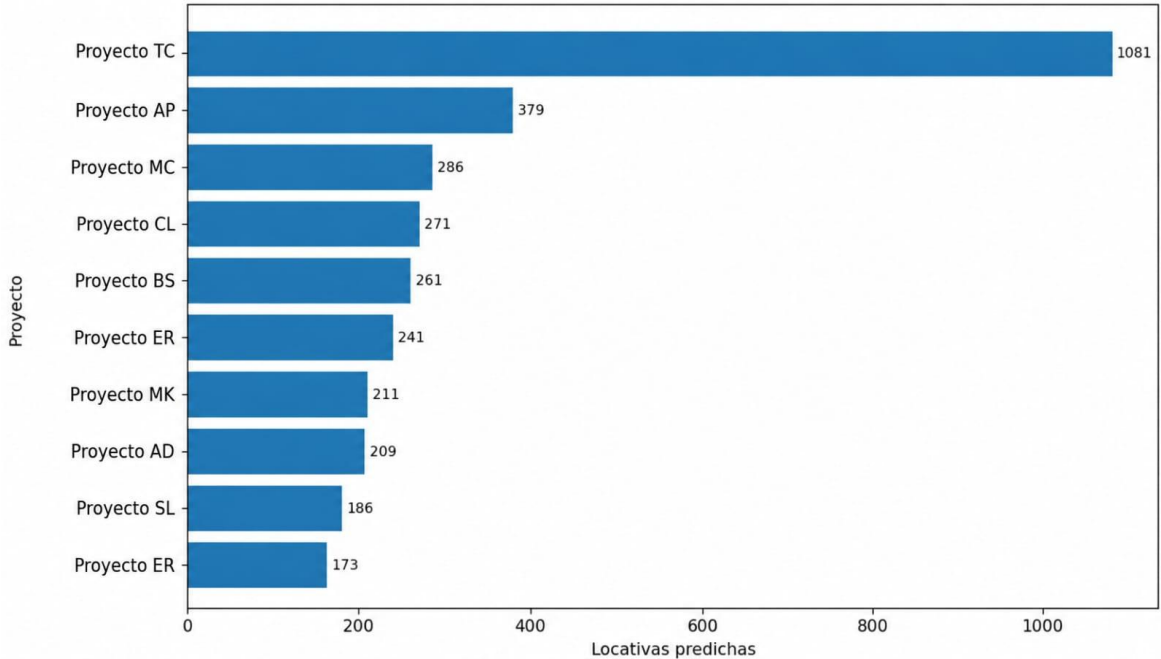
ProyectoMacro	Locativas predichas	Inmuebles expuestos	Locativas redondeadas
Proyecto TC	1.081,243896	44.396	1.081
Proyecto AP	379,360718	31.548	379
Proyecto MC	285,619354	55.704	286
Proyecto CL	271,380676	63.118	271
Proyecto BS	260,980804	55.044	261
Proyecto ER	241,492203	77.814	241
Proyecto MK	210,860031	14.542	211
Proyecto AD	208,555069	47.630	209
Proyecto SL	185,561081	15.950	186
Proyecto ER	173,378815	29.744	173

Fuente: Elaboración propia

Los resultados muestran que *Proyecto TC* es el proyecto con mayor cantidad esperada de locativas en 2026, con aproximadamente 1.081 reclamaciones estimadas. Le siguen *Proyecto AP*, con cerca de 379 locativas; *Proyecto MC*, con aproximadamente 286; *Proyecto CL*, con cerca de 271; y *Proyecto BS*, con aproximadamente 261.

Esta concentración por proyecto no debe interpretarse automáticamente como un indicador de mala calidad constructiva, ya que la predicción depende también del número de inmuebles expuestos, del volumen de entregas futuras, de las características del proyecto, del tipo de vivienda, del tipo de cubierta, del acabado y de patrones históricos observados, por lo tanto, *un proyecto con alta predicción puede reflejar una combinación de mayor exposición, mayor volumen de unidades entregadas o antecedentes de comportamiento posventa en categorías específicas*. No obstante, la identificación de proyectos con mayor carga esperada sí constituye una señal útil para la planeación, ya que estos proyectos pueden priorizarse para seguimiento preventivo, disponibilidad operativa o medidas de contención.

Ilustración 5. Proyectos con mayor predicción de locativas en 2026



Fuente: Elaboración propia en VisualStudio

5.5.Resultados por padre o capítulo

El análisis por padre o capítulo homologado permite identificar qué tipos de garantías se esperan con mayor frecuencia durante 2026. Esta vista es relevante porque no todas las reclamaciones tienen el mismo impacto operativo, técnico o económico, algunas categorías pueden requerir personal especializado, contratistas específicos o revisiones preventivas desde obra.

Tabla 13. Padres/capítulos con mayor predicción de locativas en 2026

PadreHomologado	Locativas predichas	Locativas redondeadas
CUBIERTAS	792,573975	793
MUROS	454,122009	454
HIDROSANITARIO	405,964874	406
VENTANAS VIDRIOS	379,987335	380
ELECTRICO	291,182068	291
PISOS	245,999115	246
APARATOS SANITARIOS	226,117599	226
MADERA	208,944214	209
METALICA	183,256180	183
ACABADO	161,130066	161

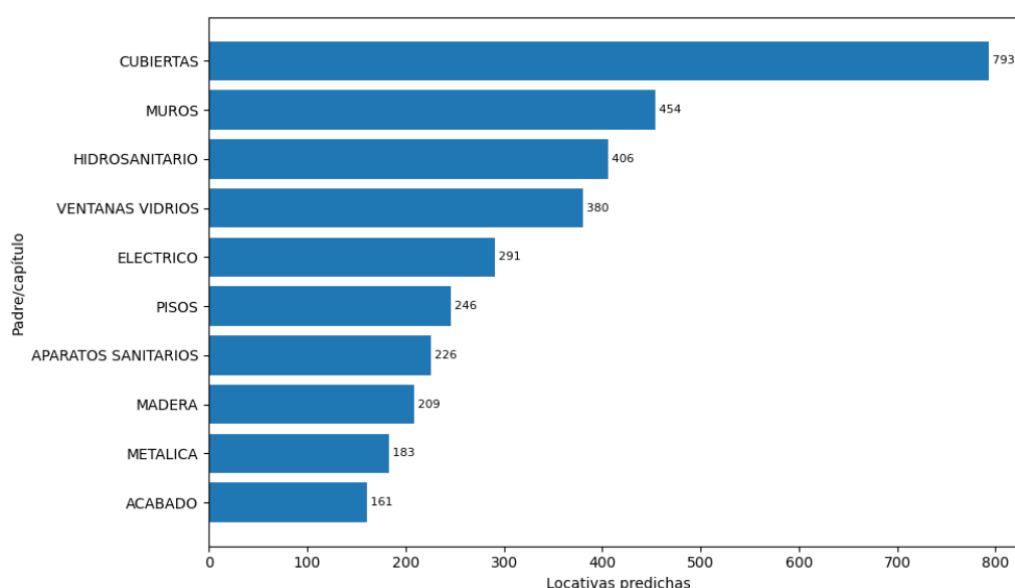
Fuente. Elaboración propia

Los resultados muestran que el padre con mayor predicción es *CUBIERTAS*, con aproximadamente 793 locativas esperadas. Le siguen *MUROS*, con cerca de 454; *HIDROSANITARIO*, con aproximadamente 406; *VENTANAS VIDRIOS*, con cerca de 380; y *ELÉCTRICO*, con aproximadamente 291.

La concentración en *CUBIERTAS* es coherente con la importancia técnica de esta categoría dentro de la posventa, especialmente en proyectos donde la exposición a condiciones ambientales, sistemas de impermeabilización o características constructivas pueden incidir en reportes relacionados con filtraciones, goteras o humedades. Sin embargo, al igual que con los proyectos, estos resultados deben interpretarse como señales predictivas y no como relaciones causales directas, y están íntimamente relacionadas a las características climáticas del Valle de Cauca.

La presencia relevante de categorías como *MUROS*, *HIDROSANITARIO*, *VENTANAS VIDRIOS* y *ELÉCTRICO* indica que la gestión futura de garantías no debe centrarse únicamente en una familia técnica, sino mantener una visión integral de las principales tipologías esperadas. Esta información puede apoyar la preparación de materiales, contratistas, protocolos de diagnóstico y estrategias de atención diferenciadas por tipo de daño.

Ilustración 6. Padres/capítulos con mayor predicción de locativas en 2026



Fuente: Elaboración propia en VisualStudio

5.6.Lectura conjunta: mes, proyecto y padre/capítulo

Una de las principales ventajas del modelo es que las predicciones pueden analizarse de manera combinada: No solo es posible conocer el total esperado de locativas para 2026, sino también identificar qué proyectos concentran determinadas categorías de daño y en qué meses podría presentarse mayor carga. Esta lectura integrada resulta especialmente valiosa para la gestión operativa. Por ejemplo, si un proyecto presenta alta predicción en *CUBIERTAS* durante los meses de mayor exposición, el área de posventa puede anticipar revisiones técnicas, coordinar disponibilidad de especialistas o preparar acciones preventivas antes de que se materialicen los reclamos. De igual manera, si un proyecto concentra predicciones en categorías como *HIDROSANITARIO* o *VENTANAS VIDRIOS*, puede requerir una estrategia de atención distinta.

La matriz proyecto × padre/capítulo permite identificar combinaciones críticas, es decir, aquellas donde coinciden alta exposición, alta predicción y una categoría técnica relevante, generando una interpretación de datos que permite pasar de una lectura general del volumen esperado a una planeación más específica y gerencial por frente de atención.

5.7.Implicaciones para la gestión posventa

Los resultados de la predicción futura tienen implicaciones directas para la gestión posventa de la constructora. En primer lugar, permiten anticipar la carga mensual de reclamaciones esperadas, lo cual facilita la planeación de recursos humanos, técnicos y logísticos, en lugar de esperar a que los reclamos se presenten, así, el área puede preparar su capacidad operativa

con base en una estimación cuantitativa. En segundo lugar, los resultados permiten priorizar proyectos. Aquellos con mayor predicción de locativas pueden ser objeto de seguimiento más cercano, revisión de antecedentes, preparación de cuadrillas o coordinación anticipada con contratistas. Esto no implica asumir que dichos proyectos presentarán necesariamente problemas de calidad, sino reconocer que concentran una mayor carga esperada de atención.

En tercer lugar, la predicción por padre o capítulo permite anticipar necesidades técnicas específicas. Si se espera una mayor proporción de reclamaciones en cubiertas, muros, hidrosanitario, ventanas o eléctrico, el área puede preparar protocolos, repuestos, personal especializado o planes de revisión preventiva relacionados con esas categorías.

En cuarto lugar, el modelo puede servir como insumo para la mejora continua: Las predicciones no solo apoyan la atención posventa, sino que también pueden retroalimentar a las áreas de construcción, calidad y diseño. La identificación de categorías recurrentes o proyectos con mayor carga esperada puede orientar revisiones técnicas, análisis de causas y acciones preventivas en proyectos futuros.

5.8. Discusión de resultados

Los resultados obtenidos muestran que el modelo predictivo permite transformar las entregas futuras en una estimación anticipada de carga posventa: Esta capacidad representa un avance frente a una gestión exclusivamente reactiva, ya que permite utilizar datos históricos y variables del proyecto para anticipar posibles escenarios de atención. La estimación de 4.247 locativas para 2026 no debe entenderse como una cifra exacta e inmodificable, sino como una aproximación esperada bajo las condiciones de entrega proyectadas y los patrones históricos aprendidos por el modelo. En este sentido, el resultado debe utilizarse como *herramienta de planeación, no como una predicción determinística*.

La distribución mensual evidencia que la carga esperada aumenta hacia el último trimestre del año. Este comportamiento se explica por la acumulación de inmuebles expuestos y por la expansión mensual de las entregas dentro de la ventana de garantía, y gracias a la identificación de este patrón es posible anticipar periodos de mayor presión operativa.

La concentración por proyecto muestra que una parte importante de la carga esperada se ubica en pocos macroproyectos, lo cual es útil para priorizar la gestión, pero debe interpretarse en conjunto con el volumen de inmuebles expuestos; los proyectos con más unidades entregadas o mayor exposición pueden presentar mayores conteos absolutos sin que ello implique necesariamente una tasa de reclamación más alta. La concentración por padre o capítulo confirma la relevancia de ciertas categorías técnicas dentro del comportamiento posventa. *CUBIERTAS, MUROS, HIDROSANITARIO, VENTANAS VIDRIOS y ELÉCTRICO* aparecen como las principales familias esperadas, lo cual ofrece insumos concretos para la preparación técnica del área.

En conjunto, los resultados demuestran que el modelo puede *apoyar la toma de decisiones al proporcionar una visión anticipada, segmentada y cuantitativa del comportamiento esperado de las garantías*. Su mayor valor no está en reemplazar el criterio técnico, sino en complementar la experiencia del negocio con evidencia derivada de datos históricos y entregas proyectadas.

5.9. Alcance de los resultados del modelo

Si bien el modelo XGBoost con variables de historial presentó el mejor desempeño relativo entre las alternativas evaluadas, las métricas obtenidas evidencian una capacidad explicativa moderada sobre la variable objetivo. En consecuencia, los resultados deben interpretarse como estimaciones orientadas a apoyar la planeación operativa y no como predicciones exactas del comportamiento futuro de las reclamaciones.

La naturaleza del fenómeno estudiado implica la existencia de múltiples factores que pueden influir en la generación de locativas y que no se encuentran disponibles en las fuentes de información utilizadas. Entre ellos se incluyen aspectos relacionados con materiales específicos, contratistas, condiciones particulares de ejecución, severidad de los daños, reincidencias, condiciones de uso por parte de los propietarios y decisiones operativas posteriores a la entrega de los inmuebles.

Adicionalmente, la variable objetivo presenta una alta proporción de observaciones con valor cero, característica habitual en problemas asociados a mantenimiento, calidad y gestión de garantías, donde la mayoría de las combinaciones posibles no generan reclamaciones durante un periodo determinado. Esta condición incrementa la complejidad del problema y limita la capacidad de los modelos para capturar completamente la variabilidad observada.

Por lo anterior, el modelo desarrollado no debe interpretarse como una herramienta para determinar con exactitud la cantidad de reclamaciones que ocurrirán en un proyecto específico, sino como un mecanismo de apoyo para identificar tendencias generales, estimar cargas esperadas de atención y facilitar la asignación anticipada de recursos dentro del proceso de posventa.

6. HERRAMIENTA WEB DE VISUALIZACIÓN

Este capítulo presenta la herramienta web desarrollada para consultar e interpretar los resultados generados por el modelo predictivo de garantías posventa. La herramienta fue concebida como un visualizador operativo de predicciones, orientado a facilitar la exploración de la carga esperada de locativas para el año 2026 por mes, proyecto y padre/capítulo técnico. A diferencia del notebook de modelado, la página web *no entrena, recalcula ni modifica el modelo predictivo*, siendo su función consumir el archivo final de resultados generado previamente por el modelo, específicamente el archivo *predicciones_futuras_xgb_historial.xlsx*, y presentar la información de manera clara, interactiva y útil para usuarios técnicos y de negocio. Esta decisión permite separar el proceso analítico de entrenamiento y predicción, realizado en Python, de la consulta operativa de resultados, realizada desde la interfaz web.

Aunque el modelo genera predicciones detalladas por múltiples combinaciones de variables, su utilidad práctica depende de que estas puedan ser consultadas de forma sencilla, filtrada y comprensible. Por esta razón, la página organiza los resultados en vistas ejecutivas, gráficos comparativos, rankings, matriz de calor y tabla de detalle.

6.1. Propósito de la herramienta

El propósito principal de la herramienta web es facilitar la consulta de las predicciones de garantías posventa para 2026, permitiendo identificar de manera rápida los meses, proyectos y padres/capítulos con mayor carga esperada de locativas, funcionando como un puente entre el modelo predictivo y la gestión operativa del negocio, ya que traduce los resultados técnicos en visualizaciones útiles para la planeación.

Desde una perspectiva ejecutiva, debemos permitir que se respondan preguntas como: cuántas locativas se esperan durante el año, cuál es el mes de mayor carga, qué proyecto concentra la mayor predicción y qué categoría técnica tiene mayor incidencia esperada. Desde una perspectiva operativa, permite explorar los resultados por filtros, revisar el comportamiento mensual de un proyecto específico, identificar combinaciones críticas entre proyecto y tipología, y consultar el detalle completo de las predicciones. Esta funcionalidad es relevante porque la predicción por sí sola no garantiza apropiación por parte del negocio, para que el modelo pueda apoyar decisiones reales sus resultados deben presentarse en un formato comprensible, consultable y alineado con las necesidades del área de posventa.

Ilustración 7. Vista general del visualizador de predicciones de garantías 2026



La figura muestra la estructura general de la herramienta web, incluyendo el panel de filtros globales, el encabezado de la página, las pestañas de navegación, el resumen ejecutivo y las primeras visualizaciones de resultados.

Fuente. Elaboración propia

6.2. Archivo de entrada y estructura de resultados

El archivo *predicciones_futuras_xgb_historial.xlsx*, generado por el flujo de modelado en Python es el insumo principal que consume la herramienta, este archivo contiene las predicciones futuras calculadas por el modelo XGBoost con historial reciente, así como diferentes hojas de resumen que facilitan la consulta de los resultados. El archivo de entrada permite consultar variables como ProyectoMacro, PeriodoExposicion, PadreHomologado, TipoProyecto, TipoVivienda, TipoAcabadoModelo, TipoCubierta, Zona, MesesDesdeEntrega, InmueblesExpuestos y PredLocativas. Estas variables conservan la estructura del modelo y permiten que el usuario explore tanto resultados agregados como registros detallados.

La página utiliza principalmente la *hoja de detalle de predicciones*, desde la cual recalcula las vistas filtradas según las selecciones del usuario, permitiendo que los filtros globales afecten todas las visualizaciones de forma consistente. Además, la herramienta presenta información derivada en forma de indicadores ejecutivos, gráficos mensuales, rankings por proyecto, rankings por padre/capítulo y matriz proyecto x tipología.

La interfaz incluye una aclaración metodológica importante: *el alcance de la página corresponde a la visualización de predicciones para 2026, sin entrenamiento ni ejecución del modelo desde la interfaz*. Su arquitectura separa el componente analítico del componente de visualización: el entrenamiento, validación y generación de predicciones se realiza previamente en el notebook del modelo, mientras que la página consume el archivo consolidado para presentar indicadores, gráficos, rankings, matriz y detalle filtrable. Esta

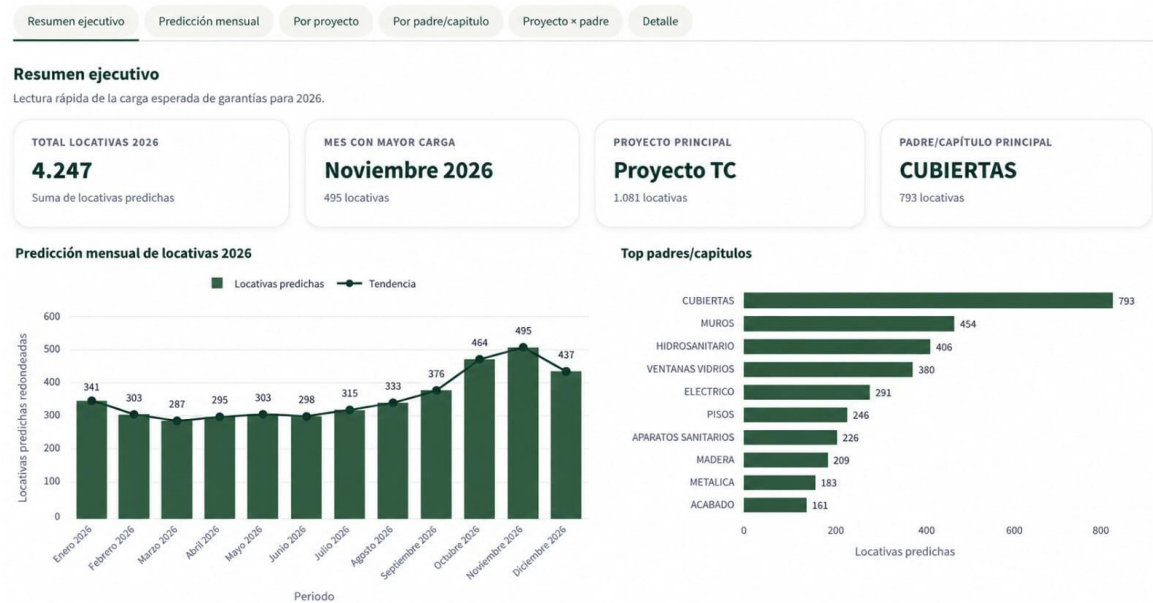
separación evita que usuarios no expertos modifiquen variables de entrada o ejecuten predicciones sin control metodológico, haciendo más estable y trazable el uso operativo de los resultados.

6.3.Vistas principales de la página web

La página se organizó en diferentes vistas o pestañas, cada una orientada a responder una necesidad particular de consulta. Esta estructura permite pasar de una lectura general a un análisis más específico, manteniendo una navegación sencilla:

La primera vista corresponde al resumen ejecutivo, donde se presentan los principales indicadores de la predicción para 2026. Esta sección incluye tarjetas con el total de locativas esperadas, el mes con mayor carga, el proyecto principal y el padre/capítulo con mayor predicción. Además, incorpora una visualización mensual de locativas y un ranking de padres/capítulos. Esta vista facilita una lectura rápida del comportamiento esperado y permite comunicar los resultados principales sin necesidad de revisar tablas detalladas. En el caso analizado, el resumen muestra una predicción total de *4.247 locativas para 2026*, el mes de *noviembre* como el de mayor carga, *Proyecto TC* como proyecto principal y *CUBIERTAS* como padre/capítulo con mayor predicción.

Ilustración 8. 6.2. Resumen ejecutivo de predicciones para 2026



Fuente. Elaboración propia

La segunda vista corresponde a la predicción mensual, en la que se visualiza la distribución de locativas predichas mes a mes durante 2026; Esta vista permite identificar tendencias

temporales y anticipar periodos de mayor demanda operativa. Aunque el informe ya presenta la predicción mensual como resultado del modelo, su inclusión en la página permite que el usuario la consulte de forma interactiva y filtrada, en caso de que se requiera.

La tercera vista corresponde al análisis por proyecto; Esta sección permite identificar los proyectos con mayor carga esperada de atención y explorar un proyecto específico, y la herramienta incluye un ranking de proyectos y un selector que permite consultar la predicción mensual y la distribución por padres/capítulos para el proyecto seleccionado. Esta funcionalidad es útil para priorizar seguimiento, revisar necesidades de capacidad operativa y preparar estrategias específicas por proyecto.

Ilustración 9. Consulta de predicciones por proyecto



Fuente: Elaboración propia

La cuarta vista corresponde al análisis por padre/capítulo, donde se presentan las tipologías técnicas con mayor cantidad esperada de locativas. Esta vista permite anticipar qué tipos de garantías podrían concentrar mayor atención durante el año, por ejemplo cubiertas, muros, hidrosanitario, ventanas vidrios o eléctrico. Aunque esta información también se resume en el tablero ejecutivo, la vista específica permite ajustar el número de categorías a mostrar y revisar el ranking de forma más detallada.

La quinta vista corresponde a la *matriz proyecto × padre/capítulo*. Esta visualización utiliza un mapa de calor para ubicar combinaciones con mayor carga esperada. Las filas representan proyectos y las columnas representan padres o capítulos técnicos; el valor de cada celda corresponde a la cantidad de locativas predichas. Esta vista es especialmente útil porque permite identificar concentraciones específicas que no siempre son evidentes en rankings

separados. Por ejemplo, un proyecto puede no ser el primero en el ranking general, pero puede presentar una concentración importante en una categoría técnica específica, de igual forma, una categoría puede tener una carga alta en el total anual, pero estar concentrada en pocos proyectos. La matriz permite reconocer estas combinaciones y orientar acciones de seguimiento más focalizadas.

Ilustración 10. Matriz proyecto por padre/capítulo



Fuente: Elaboración propia

Finalmente, la ultima incluye una vista de detalle de predicciones, que presenta una tabla con los registros filtrados. Esta tabla permite consultar los resultados a un nivel más técnico, incluyendo variables del proyecto, periodo de exposición, padre homologado, características del inmueble o proyecto, inmuebles expuestos y predicción de locativas. Además, la herramienta permite descargar el detalle filtrado en formato CSV, lo que facilita su uso posterior en análisis complementarios o reportes internos.

Ilustración 11. Detalle filtrable de predicciones

Resumen ejecutivo

Predicción mensual

Por proyecto

Por padre/capítulo

Proyecto x padre

Detalle

Detalle de predicciones

Tabla de respaldo técnico para consultar el resultado completo filtrado.

ProyectoMacro	PeriodeExposicion	PadreHomologado	TipoProyecto	TipoVivienda	TipoAcabadoModelo	TipoCubierta	Zona	MesesDesdeEntrega	InmueblesExpuestos	PredLocaliva
Proyecto ER	2026-01	CUBIERTAS	VIP	CASAS	SIN ACABADOS	FIBROCEMENTO	EL CERRITO	1	153	5.7809
Proyecto ER	2026-11	VENTANAS VIDRIOS	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	1	280	5.6476
Proyecto ER	2026-11	CUBIERTAS	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	1	280	4.9846
Proyecto ER	2026-12	VENTANAS VIDRIOS	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	1	159	4.9198
Proyecto ER	2026-02	CUBIERTAS	VIP	CASAS	SIN ACABADOS	FIBROCEMENTO	EL CERRITO	2	153	4.3141
Proyecto ER	2026-11	HIDROSANITARIO	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	1	280	4.2795
Proyecto ER	2026-10	CUBIERTAS	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	0	280	4.2744
Proyecto BS	2026-02	CUBIERTAS	VIS	APTO	SIN ACABADOS	LOSA	JAMUNDI	2	107	4.2445
Proyecto ER	2026-12	VENTANAS VIDRIOS	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	2	280	4.1051
Proyecto ER	2026-12	CUBIERTAS	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	1	159	4.0681
Proyecto TC	2026-11	MUROS	NO VIS	APTO	CON ACABADOS	LOSA	CALI	11	21	3.8261
Proyecto ER	2026-12	CUBIERTAS	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	2	280	3.7483
Proyecto BS	2026-01	CUBIERTAS	VIS	APTO	SIN ACABADOS	LOSA	JAMUNDI	1	107	3.4705
Proyecto ER	2026-12	HIDROSANITARIO	VIS	APTO	SIN ACABADOS	LOSA	ARMENIA	1	159	3.4094
Proyecto ER	2026-03	CUBIERTAS	VIP	CASAS	SIN ACABADOS	FIBROCEMENTO	EL CERRITO	3	153	3.4031

Descargar detalle filtrado en CSV

Fuente: Elaboración propia

6.4. Filtros e interactividad

Incorporamos filtros globales que recalculan todas las vistas a partir de la hoja de detalle de predicciones, como lo son año, mes, proyecto y padre/capítulo, además de una sección de filtros avanzados, orientada a ampliar la capacidad de segmentación cuando se requiera explorar escenarios más específicos. El uso de filtros permite que el usuario pase de una lectura general del año 2026 a una consulta particular por mes, proyecto o categoría técnica. Esta funcionalidad es importante porque la gestión posventa no requiere únicamente conocer el total anual esperado, sino *poder responder preguntas específicas de operación*, como qué proyectos requieren mayor seguimiento en un mes determinado o qué tipos de garantías se concentran en un proyecto específico.

La interactividad también se refleja en las vistas por proyecto y por padre/capítulo, donde el usuario puede modificar el número de elementos mostrados en los rankings. Esta característica evita saturar la visualización y permite adaptar el análisis al nivel de detalle requerido.

6.5. Uso operativo de la herramienta

Desde el punto de vista operativo, la herramienta puede apoyar diferentes actividades del área de posventa.

- ✓ Permite anticipar la carga mensual de locativas esperadas, facilitando la planeación de recursos, agendas de visita y disponibilidad de personal técnico.
- ✓ Facilita identificar proyectos con mayor carga esperada, lo que puede orientar acciones de seguimiento preventivo o preparación anticipada de contratistas.
- ✓ Simplifica el análisis de las categorías técnicas con mayor predicción, lo que puede apoyar la preparación de materiales, repuestos, cuadrillas especializadas o protocolos de diagnóstico para padres/capítulos como cubiertas, muros, hidrosanitario, ventanas vidrios o eléctrico.
- ✓ La visual de la matriz proyecto × padre/capítulo facilita la identificación de combinaciones críticas que requieren seguimiento particular, además de que su formato matricial facilita la interpretación de la información
- ✓ La vista de detalle y la posibilidad de descargar el resultado filtrado en CSV también aportan valor operativo, ya que permiten utilizar la información fuera de la página, por ejemplo en reportes internos, mesas de seguimiento o análisis específicos por equipo.

6.6. Valor aplicado para la gestión posventa

El principal valor aplicado de la herramienta consiste en convertir las predicciones del modelo en información consultable y accionable. El modelo predictivo genera estimaciones detalladas, pero la herramienta permite que esas estimaciones sean interpretadas por usuarios que no necesariamente participan en la construcción técnica del modelo. En este sentido, la página contribuye a cerrar la brecha entre ciencia de datos y operación. Los resultados dejan de estar únicamente en un notebook o archivo de salida y se convierten en una interfaz de consulta que permite explorar la información de manera visual. Esto favorece la apropiación de los resultados por parte del negocio y facilita su incorporación en procesos de planeación.

Por otra parte, se refuerza el carácter reproducible del proyecto al consumir un archivo de salida generado por el modelo, manteniendo una separación clara entre el proceso analítico y la visualización. De esta manera, futuras actualizaciones del modelo podrían generar nuevos archivos de predicción que alimenten la página sin modificar necesariamente la estructura de la interfaz.

Finalmente, la herramienta aporta a la transición de una gestión posventa reactiva hacia una gestión más anticipada, sabiendo que los negocios actuales requieren un enfoque diferenciador entre las compañías, ya que la competencia es fuerte y son los que logran hacer protagonista a la innovación y la adopción de las nuevas tecnologías quienes logran crear el distintivo que marque la preferencia por su marca o servicio, y las herramientas de la ciencia de datos nos permiten lograrlo.

6.7. Ejemplo de consulta operativa

Para ilustrar el uso de la herramienta, se puede considerar una consulta orientada a la planeación del área de posventa para el año 2026. Al ingresar a la página, el usuario observa en el resumen ejecutivo que el modelo estima aproximadamente 4.247 locativas para 2026, con noviembre como el mes de mayor carga esperada. Posteriormente, al revisar la vista por proyecto, identifica que el Proyecto TC concentra la mayor predicción de locativas. Al seleccionar este proyecto, la herramienta permite observar su comportamiento mensual esperado y la distribución de locativas por padre/capítulo. Finalmente, mediante la matriz proyecto x tipología y la tabla de detalle, el usuario puede identificar qué categorías técnicas concentran mayor carga y descargar la información filtrada en formato CSV para análisis complementarios o seguimiento interno.

Este ejemplo muestra cómo la herramienta permite pasar de una lectura ejecutiva del total esperado a una consulta operativa específica por proyecto y categoría técnica. De esta manera, el área de posventa puede anticipar periodos de mayor demanda, priorizar proyectos con mayor carga esperada y preparar recursos técnicos de acuerdo con las tipologías de garantía más relevantes.

7. CONCLUSIONES Y TRABAJOS FUTUROS

El proyecto permitió desarrollar un modelo predictivo para estimar el comportamiento de las garantías posventa en proyectos residenciales de la constructora objeto de este estudio ubicada en la ciudad de Cali, a partir de la integración de información histórica de entregas, locativas de garantía, características técnicas de los proyectos, tipo de acabado y exposición mensual de inmuebles. Con ello, se avanzó desde una lectura exclusivamente reactiva de la posventa hacia una aproximación analítica que permite anticipar cargas esperadas de reclamaciones y apoyar la planeación operativa con base en datos.

En relación con el primer objetivo específico, se logró preprocesar y estructurar información proveniente de diferentes fuentes internas y tablas auxiliares. Este proceso incluyó la depuración de registros, la homologación de macroproyectos, la estandarización de padres o capítulos técnicos, la integración de entregas con garantías y la construcción de variables relevantes para el modelo. Una decisión metodológica clave fue la construcción de la exposición mensual de inmuebles entre los meses 0 y 12 posteriores a la entrega, lo cual permitió relacionar las reclamaciones con la cantidad real de unidades susceptibles de presentar garantías.

Respecto al segundo objetivo específico, se desarrollaron y compararon diferentes modelos supervisados de regresión para estimar la cantidad esperada de locativas. Se evaluaron modelos base, modelos de ensamble, regresión de Poisson y versiones con variables de historial reciente. El modelo seleccionado fue XGBoost con historial, dado que presentó el mejor equilibrio técnico general entre las alternativas evaluadas, con el menor RMSE, el mayor R^2 y el menor MAPE sobre registros positivos.

En cuanto al tercer objetivo específico, el desempeño del modelo fue evaluado mediante métricas cuantitativas de regresión. El modelo seleccionado obtuvo un MAE de 0,111872, un RMSE de 0,465917, un R^2 de 0,235528 y un MAPE sobre registros positivos de 68,834192 %. Además, en el conjunto de prueba el total real fue de 3.328 locativas y el total predicho fue de 3.829,739258. Estos resultados muestran que el modelo mejora la línea base y permite generar estimaciones útiles para la planeación agregada, aunque debe interpretarse como una herramienta de apoyo y no como una predicción exacta de eventos individuales.

El análisis de importancia de variables evidenció que el comportamiento reciente por padre o capítulo fue el principal factor predictivo del modelo. La variable *LocativasPadreMesAnterior* tuvo la mayor importancia relativa, seguida por *ProyectoMacro*, *PadreHomologado*, *LocativasPadreUltimos3Meses* y *PromedioLocativasPadreUltimos3Meses*. Este resultado confirma que las garantías posventa presentan dependencia temporal y que el historial reciente de una categoría técnica dentro de un proyecto constituye una señal relevante para anticipar reclamaciones futuras. Además, dentro de la elaboración del modelo se tuvo una retroalimentación constante de un conocedor del negocio, lo que nos permitió reafirmar la pertinencia de los resultados, lógica y congruencia según la experiencia, y en algunos casos la

realidad, debido a que el modelo se construyó con una base que finaliza en octubre 2025 y a la fecha de escrito este documento ya se cuenta con varios meses del año 2026 para realizar una validación real de su efectividad.

La aplicación del modelo sobre las entregas futuras permitió estimar una carga esperada de aproximadamente 4.247 locativas para 2026. Los resultados mostraron una mayor concentración hacia el último trimestre del año, especialmente en noviembre. A nivel de proyecto, el Proyecto TC presentó la mayor predicción de locativas, seguido por Proyecto AP, Proyecto MC, Proyecto CL y Proyecto BS. A nivel de padre o capítulo técnico, las categorías con mayor carga esperada fueron CUBIERTAS, MUROS, HIDROSANITARIO, VENTANAS VIDRIOS y ELÉCTRICO. Estos hallazgos permiten orientar acciones de planeación, seguimiento preventivo y preparación de recursos técnicos.

La herramienta web desarrollada fortaleció el componente aplicado del proyecto, al convertir las predicciones generadas por el modelo en información visual, filtrable y accionable para el área de posventa. Su diseño permite consultar la carga esperada de garantías para 2026 por mes, proyecto y padre/capítulo, sin ejecutar nuevamente el modelo ni alterar el proceso analítico. Esta decisión fue relevante porque la predicción depende de múltiples variables técnicas, históricas, temporales y estructurales, por lo que permitir el ingreso libre de información desde la interfaz podía generar errores de captura, interpretaciones equivocadas o resultados poco confiables por parte de usuarios no expertos en el modelo o en la estructura de los datos. Por esta razón, se optó por blindar la visualización mediante el consumo de un archivo final previamente validado, haciendo la herramienta más estable, controlada y fácil de usar. En consecuencia, la página orienta al usuario sobre la información a consumir y su interpretación, evitando sesgos de malinterpretación de las instrucciones, apoyando decisiones de planeación operativa de manera comprensible, trazable y fundamentada en datos.

En conjunto, el proyecto demostró que los datos históricos de garantías pueden transformarse en un insumo útil para la toma de decisiones. La solución desarrollada no reemplaza el criterio técnico del área de posventa, pero sí aporta una base cuantitativa para anticipar cargas operativas, identificar focos de seguimiento y fortalecer la gestión preventiva. Esto representa un avance en la aplicación de ciencia de datos a un proceso tradicionalmente reactivo dentro del sector constructor.

7.1.Limitaciones del proyecto

La primera limitación del proyecto está relacionada con la calidad y disponibilidad de los datos históricos, las bases utilizadas provienen de sistemas operativos que han tenido cambios en su estructura, taxonomía y criterios de registro a lo largo del tiempo, y aunque se realizaron procesos de homologación y depuración, pueden persistir diferencias asociadas a prácticas de digitación, cambios administrativos o variaciones en la clasificación técnica de las garantías.

Una segunda limitación corresponde a la exclusión de reportes de zonas comunes, ya que estos registros fueron omitidos del modelo principal porque no cuentan con una estructura de exposición comparable a la de los inmuebles privados entregados. Para modelar zonas comunes sería necesario construir una base específica con inventario de espacios, características de equipamiento, fechas de entrega o recibo y criterios de exposición propios.

Otra limitación corresponde a las locativas sin entrega asociada. Estos casos fueron excluidos porque, al no contar con fecha de entrega, no era posible calcular los meses desde entrega ni la exposición del inmueble. Aunque esta decisión fortaleció la consistencia metodológica del modelo, también redujo el universo de datos disponible para entrenamiento.

El modelo también se encuentra limitado por la disponibilidad de variables explicativas, es decir, factores como contratistas, cuadrillas, materiales específicos, controles de calidad en obra, severidad del daño, costo de reparación, reincidencias, percepción del cliente o condiciones de uso del inmueble, que no estuvieron disponibles de forma estructurada para esta versión. Estas variables podrían explicar parte de la variabilidad no capturada por el modelo, y puede enriquecer el mismo a nivel de exactitud. Sin embargo, es importante considerar que en dicho escenario puede darse a lugar una sobre especificación del modelo que altere su aplicabilidad general.

Las variables climáticas fueron exploradas como fuente contextual, especialmente por su posible relación con reclamaciones asociadas a cubiertas, humedades, filtraciones y goteras, pero lastimosamente, aunque se consideraron en un inicio como una variable clave en la predicción, no se incorporaron como predictores principales del modelo final debido a limitaciones de cobertura, completitud y consistencia temporal. Por tanto, su integración queda como una oportunidad de fortalecimiento para futuras versiones.

Finalmente, el modelo estima cantidades esperadas de locativas en escenarios agregados. No está diseñado para predecir con certeza la ocurrencia individual de una garantía ni para reemplazar la evaluación técnica de los profesionales de posventa. Sus resultados deben interpretarse como señales de apoyo para la planeación, priorización y seguimiento preventivo.

7.2.Trabajos futuros

Como trabajo futuro, se recomienda implementar un proceso periódico de actualización y reentrenamiento del modelo con nuevos registros de entregas, garantías y características de proyectos, esto permitiría que el modelo incorpore cambios recientes en procesos constructivos, materiales, proveedores, tipologías de proyectos y prácticas de atención posventa, que, aunque no son variables incluidas en el modelo, su variabilidad repercute directamente en el comportamiento de las reclamaciones de los proyectos.

Alineado al punto anterior, también se recomienda ampliar el conjunto de variables disponibles, siendo la incorporación de información sobre contratistas, cuadrillas, materiales,

costos de reparación, severidad del daño, tiempos de atención, reincidencias y resultados de inspecciones de calidad fuentes valiosas que podría mejorar la capacidad explicativa del modelo y permitir análisis más profundos sobre las causas asociadas a las reclamaciones. Aún así, es importante considerar que en ese punto puede ser necesario descartar algunas variables para evitar la sobre especificación del modelo, ya que al ser muy numerosas puede reducirse la cobertura de su predicción. No obstante, ese ejercicio puede permitir determinar cuáles características dan lugar a la mayor exposición a generar garantías, y a nivel del negocio evaluar el ajuste de las mismas para reducir la incidencia de los reclamos.

Otra línea de trabajo futuro consiste en fortalecer la integración de variables climáticas, para ello sería necesario consolidar una fuente climática más robusta, con datos confiables de precipitación, humedad, temperatura y eventos extremos a nivel local. Esta integración sería especialmente relevante para analizar categorías sensibles al agua, como cubiertas, humedades, filtraciones y goteras.

Se tuvo conocimiento por parte de quienes hacen parte de la empresa, que actualmente a nivel corporativo las empresas bajo el nombre de la constructora a nivel regional, se han integrado a nivel nacional, lo que genera la apertura de integrar a este modelo numerosas nuevas regiones del país, como Bogotá, Ibagué, Rionegro, Cartagena, Santamarta, Barranquilla entre otros. En este sentido, puede ser muy enriquecedor escalar la aplicabilidad del modelo a nivel Nacional, generando seguramente comportamientos diferenciadores entre regiones, y permitiendo evaluar oportunidades de mejora en todos los procesos de la compañía que desembocan en las garantías con una mayor contundencia.

En términos operativos, se recomienda avanzar hacia una integración más directa con los sistemas internos de la empresa. En futuras versiones, el modelo podría conectarse con fuentes actualizadas de entregas, garantías y especificaciones de proyectos, reduciendo la necesidad de cargas manuales y facilitando la actualización continua de resultados.

Finalmente, se recomienda fortalecer el gobierno de datos asociado al proceso de garantías posventa. La calidad del modelo depende directamente de la consistencia de los registros, por lo que resulta fundamental mantener catálogos homologados, criterios claros de clasificación técnica, trazabilidad de fechas, completitud de campos y reglas estandarizadas de registro. Esta mejora no solo beneficiaría futuras versiones del modelo, sino también la gestión diaria del proceso de posventa.

BIBLIOGRAFÍA

- [1] M. &. Company, «Reinventing construction through a productivity revolution,» [En línea]. Available: <https://www.mckinsey.com/capabilities/operations/our-insights/reinventing-construction-through-a-productivity-revolution>. [Último acceso: 29 05 2025].
- [2] Congreso de Colombia, «Ley 1480 de 2011, por medio de la cual se expide el Estatuto del Consumidor y se dictan otras disposiciones,» *Diario Oficial No. 48.220*, 12 de octubre de 2011.
- [3] Presidencia de la República de Colombia, «Decreto 735 de 2013, por el cual se reglamenta la efectividad de la garantía prevista en los artículos 7° y siguientes de la Ley 1480 de 2011,» *Diario Oficial No. 48.764*, 17 de abril de 2013.
- [4] N. Forcada, M. Macarulla, A. Fuertes, M. Casals, M. Gangolells, and X. Roca, «Post-handover housing defects: Sources and origins,» *Journal of Performance of Constructed Facilities*, vol. 27, nº 6, pp. 756-762, 2013.
- [5] E. Plebankiewicz, K. Zima, and D. Wieczorek, «Analysis of defects in residential buildings reported during the warranty period,» *Applied Sciences*, vol. 10, nº 17, p. Art. 6123, 2020.
- [6] S. D. Datta, M. Islam, H. Rahman Sobuz, S. A. y M. Kar, «Heliyon,» 15 03 2024. [En línea]. Available: <https://www.sciencedirect.com/science/article/pii/S2405844024029190>. [Último acceso: 1 02 2026].
- [7] A. C. Cameron and P. K. Trivedi, «Regression Analysis of Count Data,» *U.K.: Cambridge University Press*, vol. 2nd ed. Cambridge, 2013.
- [8] P. M. a. J. A. Nelder, «Generalized Linear Models,» *U.K.: Chapman and Hall*, vol. 2nd ed. London, 1989.
- [9] J. M. Hilbe, «Negative Binomial Regression,» *U.K.: Cambridge University Press*, vol. 2nd ed. Cambridge, 2011.
- [10] T. Chen and C. Guestrin, «“XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,» San Francisco, CA, USA, 2016, pp. 785-794.

- [11] L. Grinsztajn, E. Oyallon, and G. Varoquaux, «Why do tree-based models still outperform deep learning on typical tabular data?,» *Advances in Neural Information Processing Systems*, vol. 35, 2022.
- [12] R. Wirth and J. Hipp, «CRISP-DM: Towards a standard process model for data mining,» *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000.
- [13] P. Chapman et al, «CRISP-DM 1.0: Step-by-step data mining guide,» *SPSS Inc*, 2000.
- [14] C. Schröer, F. Kruse, and J. M. Gómez, «A systematic literature review on applying CRISP-DM process model,» *Procedia Computer Science*, vol. 181, pp. 526-534, 2021.
- [15] J. E. Pinzón Serpa et al, «Estudio y análisis de los factores que generan reclamaciones de posventa en la construcción de vivienda multifamiliar,» Universidad Santo Tomás, 2015. [En línea]. Available: <https://hdl.handle.net/11634/881>. [Último acceso: 30 05 2025].
- [16] M. A. Correa Manrique et al, «Lean Construction Strategies Supported by Artificial Intelligence Techniques for Construction Project Management—A Review,» 2024. [En línea]. Available: https://www.researchgate.net/publication/378539646_Lean_Construction_Strategies_Supported_by_Artificial_Intelligence_Techniques_for_Construction_Project_Management-A_Review.. [Último acceso: 29 05 2025].
- [17] R. J. Hyndman and A. B. Koehler, «Another look at measures of forecast accuracy,» *International Journal of Forecasting*, vol. 22, nº 4, pp. 679-688, 2006.