

Análisis del Sector Inmobiliario mediante Agentes Inteligentes

y Políticas Públicas Integradas



Estudiante:

Profesores:

Asignatura: 22.536 - Trabajo final de grado

Estudios: Grado de Ciencia de Datos Aplicada

Resumen

Este trabajo presenta una adaptación y aplicación del modelo basado en agentes PolicySpace2 (PS2) para analizar el mercado inmobiliario español a nivel municipal y evaluar el impacto de diversas políticas públicas. El proyecto se centra en la creación de una robusta infraestructura de datos mediante un extenso proceso de Extracción, Transformación y Carga (ETL) de fuentes oficiales españolas, incluyendo datos demográficos, económicos, catastrales y fiscales. Se detalla la metodología de integración de estos datos, la consolidación de una base de datos, y el desarrollo de modelos predictivos de precios de vivienda.

PolicySpace2, originado en Brasil con aplicaciones documentadas desde aproximadamente 2020 y publicaciones clave en 2021 y 2022, ha sido validado empíricamente en 46 regiones metropolitanas brasileñas. Se reconoce especialmente la dedicación y contribución fundamental de Bernardo Alves Furtado al desarrollo y la conceptualización de este modelo en dicho campo. Ha demostrado su capacidad para replicar indicadores macroeconómicos como el PIB, la inflación, el desempleo y el coeficiente de Gini, sirviendo como herramienta para simular y comparar mecanismos de inversión pública orientados a hogares de bajos ingresos.

La adaptación española busca replicar esta capacidad analítica, permitiendo la simulación de escenarios de políticas como vales de alquiler, ayudas monetarias y la gestión del stock de viviendas, con el fin de ofrecer información valiosa para la toma de decisiones en el complejo contexto del mercado inmobiliario español. Se presentan resultados exploratorios del proceso ETL, del modelo predictivo de precios, y de simulaciones iniciales del modelo PolicySpace2 adaptado, junto con una discusión de las contribuciones a la infraestructura de datos y las líneas de trabajo futuras.

Abstract

This work presents an adaptation and application of the agent-based model PolicySpace2 (PS2) to analyze the Spanish housing market at the municipal level and evaluate the impact of various public policies. The project focuses on creating a robust data infrastructure through an extensive Extract, Transform, Load (ETL) process from official Spanish sources, including demographic, economic, cadastral, and fiscal data. The methodology for integrating this data, consolidating a database, and developing housing price prediction models is detailed.

PolicySpace2, originating in Brazil with documented applications since approximately 2020 and key publications in 2021 and 2022, has been empirically validated in 46 Brazilian metropolitan regions. Special recognition is given to the dedication and fundamental contribution of Bernardo Alves Furtado to the development and conceptualization of this model in the field. It has demonstrated its ability to replicate macroeconomic indicators such as GDP, inflation, unemployment, and the Gini coefficient, serving as a tool to simulate and compare public investment mechanisms aimed at low-income households.

The Spanish adaptation seeks to replicate this analytical capability, allowing the simulation of policy scenarios such as rental vouchers, monetary aid, and housing stock management, in order to offer valuable information for decision-making in the complex context of the Spanish housing market. Exploratory results from the ETL process, the price prediction model, and initial simulations of the adapted PolicySpace2 model are presented, along with a discussion of the contributions to the data infrastructure and future lines of work.



Figura 1: Esta obra está sujeta a una licencia de Reconocimiento -NoComercial- SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo: Análisis del Sector Inmobiliario mediante Agentes Inteligentes y Políticas Públicas Integradas

Nombre del autor:

Nombre del director/a:

Nombre del PRA:

Fecha de entrega (mm/aaaa): 03/06/2025

Titulación o programa: Grado de Ciencia de Datos Aplicada

Área del Trabajo Final: Ciencia de Datos Aplicada a Políticas Públicas de Vivienda

Idioma del trabajo: Castellano

Palabras clave: Modelos Basados en Agentes, Inteligencia Artificial, Vivienda.

Resumen del Trabajo: Este estudio adapta el modelo PolicySpace2 (PS2) para analizar el mercado inmobiliario español a nivel municipal y estudiar el impacto de políticas públicas. Primero, se crea una infraestructura de datos mediante un proceso ETL, integrando fuentes oficiales españolas. A partir de estos datos, se desarrollan modelos predictivos para los precios de vivienda. El modelo PS2, originario de Brasil, ha demostrado replicar indicadores macroeconómicos y simula políticas clave. Se presentan resultados iniciales y se exploran futuras líneas de trabajo.

Abstract: This study adapts the PolicySpace2 (PS2) model to analyze the Spanish real estate market at the municipal level and examine the impact of public policies. First, a data infrastructure is created through an ETL process, integrating official Spanish sources. Based on these data, predictive models for housing prices are developed. The PS2 model, originally from Brazil, has demonstrated the ability to replicate macroeconomic indicators and simulate key policies. Initial results are presented, and future lines of work are explored.

Análisis del Sector Inmobiliario mediante Agentes Inteligentes y Políticas Públicas Integradas

Enero 2025

Índice

1. Introducción y Contexto	7
1.1. Revisión breve del problema central	7
1.2. Justificación de los aspectos técnico y ético/legal/social	8
2. Objetivos y Problemática a Resolver	8
2.1. Objetivos generales	8
2.2. Objetivos específicos	8
2.3. Indicadores clave de consecución	8
2.4. Problemática a resolver	9
3. Aspecto Ético/Legal/Social/Organizativo	9
3.1. Implicaciones éticas y legales del uso de datos personales	9
3.2. Impacto social de los problemas de vivienda	9
3.3. Políticas y normativas clave	9
4. Estado del Arte	9
4.1. Enfoques Basados en Datos e Inteligencia Artificial en el Sector Inmobiliario	10
4.2. Simulaciones Basadas en Agentes (ABM) en el Mercado Inmobiliario	10
4.3. Evaluación de Políticas Públicas con Econometría	11
4.4. Estudios Académicos Relevantes y Análisis Geoespacial	11
5. Solución Propuesta y Metodología	12
5.1. Solución Propuesta	12
5.2. Fuentes de Datos	14
5.3. Analítica	15
5.4. Algorítmica	16
5.5. Arquitectura de la Solución	18
5.6. Visualización	20
5.7. Ética y Legalidad	20
5.8. Resumen Metodológico	20
5.9. Estructura de Datos de Referencia y Componentes Iniciales del Proyecto PolicySpace2_Spanish_data .	20
5.9.1. Tablas de Equivalencias y Categorización de Datos	20
5.9.2. Base de Datos Inicial y Dashboard	21
5.10. Proceso de Extracción, Transformación y Carga (ETL) de Datos para España	21
5.10.1. Reducción Progresiva del Conjunto de Datos Municipal	21
5.10.2. Procesamiento de Cifras de Población por Municipio (Versión Completa con Formato Ancho) .	22
5.10.3. Procesamiento de Datos de Mortalidad por Comunidad Autónoma y Sexo	22
5.10.4. Estimación de la Proporción de Población Urbana Municipal	22
5.10.5. Procesamiento del Número de Empresas por Municipio y Actividad Principal	23
5.10.6. Generación de Estimaciones de Población Municipal Anual (Formato Ancho)	23
5.10.7. Procesamiento de Indicadores de Fecundidad por Provincia y Comunidad Autónoma	23
5.10.8. Procesamiento de Datos de Tasas de Interés	23
5.10.9. Procesamiento de Datos de Nivel Educativo por Comunidad Autónoma	24
5.10.10. Procesamiento de Datos de Participación en los Ingresos del Estado (PIE)	24
5.10.11. Procesamiento de Datos de Tamaño Medio de los Hogares por Comunidad Autónoma	24

5.10.12. Análisis Exploratorio Visual de Datos de Entrada	25
5.11. Adaptación del Modelo PolicySpace2 y Componentes Adicionales	36
5.11.1. Adecuación del Modelo y Base de Datos Mejorada (PolicySpace2_Spain_new_ETL)	36
5.11.2. Integración de Datos Catastrales y Generación de Viviendas	37
5.11.3. Modelo de Precios de Vivienda (proyecto_modelo_1)	41
5.11.4. Resultados de Simulación y Documentación Adicional	42
6. Planificación y calendario	42
7. Análisis y Gestión de Riesgos	43
7.1. Identificación de Riesgos	43
7.2. Evaluación de Probabilidad e Impacto	44
7.3. Plan de Gestión de Riesgos	44
8. Herramientas y Fuentes para la Construcción de la Base del Proyecto	44
8.1. Connected Papers	44
8.2. Mendeley	45
8.3. Google Scholar	45
8.4. Kaggle	45
8.5. VLex	45
8.6. GitHub	46
8.7. Fuentes de Datos Primarias y Organismos Oficiales	46
8.8. Entorno de Desarrollo y Lenguajes de Programación	46
8.9. Sistemas de Gestión de Bases de Datos	47
8.10. Herramientas de Visualización, BI y Documentación	47
8.11. Herramientas de Gestión de Proyectos y Colaboración	47
9. Implicaciones para el Diseño del Proyecto	47
9.1. Ética y Legalidad	47
9.2. Limitaciones Técnicas	47
9.3. Líneas de Acción	48
10. Resultados del Proyecto y Discusión	48
10.1. Contexto Analítico y Hallazgos de la Literatura Previa	48
10.1.1. Riesgos Éticos/Legales frente a Beneficios Sociales de la IA en Vivienda	48
10.1.2. Comparativa de Técnicas de ML Aplicadas al Sector Vivienda	48
10.1.3. Hallazgos Clave de la Literatura	48
10.2. Resultados del Modelo Predictivo de Precios de Vivienda	49
10.2.1. Rendimiento y Comparación de Modelos	49
10.2.2. Análisis de Variables Más Importantes	50
10.2.3. Validación del Modelo	52
10.3. Resultados de las Simulaciones con PolicySpace2 (Adaptación Española)	53
10.3.1. Escenario Base: Tendencias Agregadas y Dinámicas Municipales	53
Introducción al Escenario Base de Simulación	53
10.3.2. Impacto de Políticas Socioeconómicas Simuladas	60
10.3.3. Discusión sobre la Variabilidad y Calibración del Modelo	62
10.4. Contribuciones a la Infraestructura de Datos	63
10.4.1. Consolidación de la Base de Datos Mejorada	63
10.4.2. Integración de Datos Catastrales y su Impacto Potencial	63
Recursos Digitales y Demostraciones del Proyecto	64
11. Conclusiones y trabajos futuros	65
Glosario	67

Índice Visual

Objetivos y Problemática a Resolver: Definición de los objetivos del proyecto y los problemas centrales que se buscan abordar.

Resumen

Se establecen los objetivos generales y específicos del proyecto, así como los indicadores clave para medir el éxito de las soluciones propuestas, enfocándose en la accesibilidad a la vivienda, la gentrificación y la desigualdad económica.

Aspecto Ético/Legal/Social/Organizativo: Análisis de las implicaciones éticas, legales y sociales del proyecto.

Resumen

Este apartado aborda los retos éticos y legales relacionados con el uso de datos personales y la implementación de IA en el mercado inmobiliario, además del impacto social de fenómenos como la gentrificación y la necesidad de regulaciones adecuadas.

Aspectos Técnicos a Tener en Cuenta: Descripción de las herramientas y metodologías técnicas utilizadas en el análisis.

Resumen

Se detallan los modelos avanzados de Machine Learning utilizados, así como metodologías complementarias como el análisis de series temporales, clustering y simulaciones basadas en agentes, además de la evaluación de políticas públicas con econometría.

Solución Propuesta y Metodología: Presentación de la solución integral y la metodología adoptada.

Resumen

Se propone el uso del modelo PolicySpace2 basado en agentes para simular las dinámicas del mercado inmobiliario y evaluar el impacto de diversas políticas públicas, integrando datos demográficos, económicos, geográficos y fiscales.

Resultados y Comparativas en el Sector de la Vivienda: Presentación de los hallazgos y comparaciones con otras técnicas y estudios.

Resumen

Se exponen las comparativas de riesgos éticos frente a beneficios sociales de la IA en vivienda, técnicas de Machine Learning aplicadas al sector, y los hallazgos clave que destacan el impacto técnico-social y la regulación basada en datos.

Implicaciones para el Diseño del Proyecto: Consideraciones finales para el diseño y ejecución del proyecto.

Resumen

Se discuten las implicaciones éticas y legales, las limitaciones técnicas y las líneas de acción necesarias para asegurar la transparencia, eficiencia y cumplimiento normativo en el diseño del proyecto.

1. Introducción y Contexto

Este documento explora conceptos clave relacionados con el análisis y la gestión del mercado inmobiliario, destacando términos fundamentales como **Inteligencia Artificial (IA)**, **Big Data** y **ETL (Extract, Transform, Load)**. La **Inteligencia Artificial (IA)** permite desarrollar sistemas que simulan habilidades humanas para el análisis y la toma de decisiones, mientras que el **Big Data** proporciona herramientas avanzadas para gestionar y analizar grandes volúmenes de información. Además, procesos como **ETL (Extract, Transform, Load)** facilitan la extracción, transformación y carga de datos desde diversas fuentes hacia un almacenamiento unificado, lo que resulta esencial para proyectos de esta magnitud.

Otro concepto clave a entender en este análisis del mercado inmobiliario es el uso de **Agentes**, modelos computacionales que simulan el comportamiento de actores como compradores, vendedores y reguladores. Estos **ABM** interactúan en un entorno virtual, modelando dinámicas complejas como la formación de precios y las decisiones de localización. Este enfoque, integrado en simulaciones, permite evaluar el impacto de, como la regulación de alquileres o los subsidios a la vivienda, facilitando el diseño de soluciones más efectivas y equitativas.

Por lo tanto, el uso de **Agentes** no solo permite analizar el efecto directo de las políticas públicas sobre el mercado inmobiliario, como la reducción de precios o el incremento de la disponibilidad de vivienda, sino también podemos evaluar su impacto en los negocios locales que dependen del turismo. Por ejemplo, restricciones severas al alquiler turístico pueden provocar una disminución en el flujo de turistas que afecta directamente a restaurantes, tiendas y otros servicios locales, reduciendo sus ingresos y la generación de empleo. Estas simulaciones permiten modelar cómo estas dinámicas económicas interactúan con el mercado inmobiliario, identificando posibles compensaciones entre el acceso a la vivienda y el sustento de las economías locales, facilitando así el diseño de políticas más equilibradas y sostenibles.

1.1. Revisión breve del problema central

El mercado inmobiliario español ha experimentado transformaciones significativas en las últimas décadas debido a factores económicos, políticos y sociales. La crisis financiera de 2008 y el estallido de la burbuja inmobiliaria provocaron una caída en los precios y alquileres de la vivienda hasta 2013, cuando comenzaron a recuperarse junto con el crecimiento económico [13]. En 2024, a pesar del crecimiento sostenido del 4,3% interanual en los precios de la vivienda nueva, persisten problemas de accesibilidad debido a la oferta limitada y al aumento de los costes de construcción. Además, la proliferación de alquileres turísticos, como los ofertados en plataformas como Airbnb, ha contribuido al aumento de los precios de los alquileres residenciales y a la reducción de la oferta para residentes locales, exacerbando las dificultades para encontrar viviendas asequibles[12].

Las plataformas de alquiler vacacional como Airbnb han introducido una nueva dimensión en el mercado, generando oportunidades y retos. Por un lado, contribuyen a la economía colaborativa, pero también generan efectos negativos como la **Gentrificación** y el desplazamiento de residentes [2, 12].

La IA emerge como una herramienta crítica para analizar y abordar estas problemáticas. Su capacidad para procesar grandes volúmenes de datos y predecir tendencias permite el diseño de políticas públicas más efectivas, mejorando la accesibilidad a la vivienda y promoviendo la equidad [8].

1.2. Justificación de los aspectos técnico y ético/legal/social

El análisis socioeconómico del mercado inmobiliario es crucial para entender las fuerzas que determinan la oferta, la demanda y los precios de la vivienda. Entre los factores clave se encuentran:

- **Políticas públicas:** Los esfuerzos legislativos para regular los alquileres vacacionales y garantizar el acceso a la vivienda han mostrado resultados mixtos [10].
- **Impacto de la Inteligencia Artificial (IA):** El uso de modelos de *machine learning* puede revelar patrones ocultos en los datos del mercado inmobiliario, ayudando a formular soluciones basadas en evidencia [11].

Los aspectos éticos también son fundamentales. La protección de los datos personales, regulada por el RGPD, y la necesidad de evitar la discriminación algorítmica deben ser consideraciones prioritarias al implementar herramientas de IA [9].

2. Objetivos y Problemática a Resolver

El presente trabajo tiene como objetivo abordar las dinámicas complejas del mercado inmobiliario y su interacción con las políticas públicas, utilizando modelos basados en agentes e inteligencia artificial. A continuación, se detallan los objetivos generales y específicos que guían este proyecto, junto con los indicadores clave que medirán el éxito de las soluciones propuestas.

2.1. Objetivos generales

- Desarrollar un proceso ETL (Extracción, Transformación y Carga) robusto que garantice la disponibilidad, calidad y actualización continua de los datos necesarios para el análisis del mercado inmobiliario.
- Analizar las dinámicas del mercado inmobiliario mediante modelos de simulación basados en agentes, reproduciendo el comportamiento de los actores clave y sus interacciones.
- Evaluar el impacto de políticas públicas integradas en la accesibilidad y equidad del mercado inmobiliario, considerando múltiples ámbitos (vivienda, trabajo, crédito, territorio).
- Proponer soluciones innovadoras apoyadas en inteligencia artificial para mitigar problemas como la gentrificación y la desigualdad en el acceso a la vivienda.

2.2. Objetivos específicos

- Diseñar un modelo de simulación que represente actores clave del mercado, como compradores, vendedores y reguladores.
- Implementar técnicas de machine learning para predecir tendencias del mercado inmobiliario.
- Evaluar políticas públicas mediante simulaciones y análisis comparativos.
- Crear visualizaciones e informes para comunicar los resultados y su impacto social.

2.3. Indicadores clave de consecución

Para evaluar el éxito del proyecto, se definirán los siguientes indicadores clave:

- **Precisión de los modelos predictivos:** Evaluada mediante métricas como el RMSE y el R².
- **Impacto en la equidad:** Medido a través de índices como el índice de Gini o cambios en la accesibilidad a la vivienda.
- **Eficiencia de las políticas públicas:** Análisis coste-beneficio de las políticas simuladas.

2.4. Problemática a resolver

El mercado inmobiliario enfrenta desafíos significativos, como la falta de accesibilidad a la vivienda, la gentrificación y la desigualdad económica. Este proyecto busca abordar estos problemas mediante la integración de políticas públicas efectivas y tecnologías avanzadas, garantizando un impacto social positivo y sostenible.

3. Aspecto Ético/Legal/Social/Organizativo

3.1. Implicaciones éticas y legales del uso de datos personales

El tratamiento de datos personales para predecir precios inmobiliarios plantea retos éticos y legales. La aplicación del RGPD asegura que las prácticas de recolección y análisis de datos cumplan con principios como la minimización y la transparencia [4]. Sin embargo, el uso de IA también puede amplificar sesgos preexistentes si los algoritmos no se entrenan adecuadamente [3].

3.2. Impacto social de los problemas de vivienda

Las plataformas de alquiler vacacional han exacerbado problemas como la gentrificación y la desigualdad. En ciudades como Barcelona y Lisboa, Airbnb ha aumentado los precios de los alquileres y ha reducido la disponibilidad de vivienda para residentes permanentes [13, 2]. Estos efectos resaltan la necesidad de una regulación más estricta para equilibrar los beneficios económicos del turismo con el bienestar de las comunidades locales [10].

La IA puede ser una herramienta útil para monitorear estas dinámicas y evaluar el impacto de las políticas implementadas. Por ejemplo, modelos predictivos pueden identificar áreas vulnerables a la gentrificación antes de que se produzcan desplazamientos masivos [11].

3.3. Políticas y normativas clave

Las regulaciones existentes varían significativamente entre ciudades. En Berlín, las restricciones sobre los alquileres vacacionales han reducido su impacto negativo, mientras que en otras ciudades como Barcelona, la implementación y el cumplimiento de normativas han sido más complejos [10].

A nivel europeo, el Reglamento de Inteligencia Artificial establece directrices para garantizar un uso ético y transparente de esta tecnología en sectores críticos, aplicables al estudio del sector inmobiliario [9]. Este marco puede servir de base para el desarrollo de soluciones que integren IA de manera responsable.

4. Estado del Arte

A continuación se describen las principales estrategias y enfoques existentes para el análisis del mercado inmobiliario y la evaluación de políticas de vivienda, situando el presente trabajo en el contexto de la literatura y experiencias previas. Se revisan tanto las metodologías basadas en datos masivos e inteligencia artificial, como los modelos de simulación por agentes y las técnicas econométricas clásicas empleadas para medir el impacto de políticas. Finalmente, se mencionan estudios académicos recientes que abordan problemas similares al tratado en este proyecto.

4.1. Enfoques Basados en Datos e Inteligencia Artificial en el Sector Inmobiliario

En los últimos años ha proliferado el uso de técnicas de [Ciencia de Datos \(DS\)](#) y [Machine Learning \(ML\)](#) para estudiar el mercado inmobiliario. Gracias a la disponibilidad de grandes bases de datos (por ejemplo, portales inmobiliarios con millones de anuncios, registros catastrales, datos censales, etc.), los investigadores han aplicado algoritmos de IA para detectar patrones y realizar predicciones. Un ejemplo común es la predicción de precios de vivienda a partir de características del inmueble y del barrio: modelos de regresión, árboles de decisión y redes neuronales han logrado estimaciones cada vez más precisas del valor de mercado de propiedades, superando en ocasiones a tasaciones tradicionales [6, 1]. Asimismo, se han empleado técnicas de clustering para segmentar mercados locales o identificar “puntos calientes” de gentrificación incipiente, analizando tendencias de precios y rentas.

La IA también facilita incorporar fuentes de datos heterogéneas en el análisis inmobiliario. Por ejemplo, combinar datos geoespaciales (distancias a servicios, accesibilidad al transporte público), datos textuales (descripciones de anuncios) e incluso imágenes (fotografías de viviendas analizadas con visión artificial) en modelos integrados. Estos enfoques de [Big Data](#) inmobiliario proporcionan una visión granular y en tiempo real de las dinámicas del mercado, en contraste con estadísticas oficiales que suelen tener menor resolución espacial o temporal.

Sin embargo, los modelos puramente predictivos tienen limitaciones cuando se trata de evaluar políticas públicas. Pueden predecir qué pasará si continúan las tendencias actuales, pero no tan fácilmente simular escenarios contrafactuals donde interviene una nueva política. Aquí es donde técnicas complementarias, como la simulación basada en agentes, resultan valiosas. No obstante, conviene señalar que la literatura ha explorado enfoques híbridos: por ejemplo, usar predicciones de modelos ML como insumo o referencia para calibrar simulaciones, o viceversa, emplear simulaciones para generar datos sintéticos con los que entrenar modelos de IA. El presente trabajo se inspira en esta complementariedad, combinando predicción y simulación.

4.2. Simulaciones Basadas en Agentes (ABM) en el Mercado Inmobiliario

Las simulaciones basadas en agentes se han consolidado como una poderosa herramienta para modelar el comportamiento de múltiples actores en mercados complejos, incluyendo el inmobiliario. En un [ABM](#), cada agente (sea un individuo, hogar, empresa, etc.) toma decisiones de acuerdo con ciertas reglas o estrategias, interactuando con otros agentes y con su entorno. Esta aproximación bottom-up permite que emergan dinámicas macroeconómicas a partir de comportamientos individuales y sus interacciones.

En el ámbito de la vivienda, modelos ABM como PolicySpace y PolicySpace2 [5] han sido desarrollados para estudiar cómo distintos agentes (hogares, promotores, bancos, gobiernos locales, etc.) co-evolucionan y cómo políticas específicas influyen en sus decisiones. Por ejemplo, un modelo puede incorporar agentes “hogar” que deciden mudarse o comprar vivienda en función de su situación económica, precios y expectativas, mientras agentes “constructoras” deciden cuánto construir en función de licencias disponibles y rentabilidad esperada. Las interacciones de oferta y demanda en el mercado inmobiliario, junto con factores laborales (empleo e ingresos de los hogares) y financieros (crédito hipotecario ofrecido por bancos), generan resultados agregados como la evolución de precios de venta y alquiler, tasas de propiedad vs. alquiler, segregación residencial, etc.

Este enfoque integrado permite experimentar virtualmente con políticas: ¿Qué ocurre si se introducen vales de alquiler para familias de bajos ingresos? ¿Y si un ayuntamiento impone un límite al número de pisos turísticos por barrio, o si promueve la construcción de vivienda pública? En un ABM, podemos implementar estas reglas y observar cómo reaccionan los agentes y qué métricas de resultado cambian (producción de viviendas, consumo de las familias, desigualdad en riqueza o satisfacción de necesidades habitacionales, entre otras). La capacidad de incorporar comportamientos adaptativos (los agentes pueden aprender o cambiar sus estrategias con el tiempo) brinda aún más realismo frente a métodos estáticos.

Cabe destacar que, a diferencia de los modelos econométricos, las simulaciones por agentes no requieren suponer a priori equilibrios de mercado o relaciones lineales simples; pueden captar procesos fuera de equilibrio, retroalimentaciones no lineales y shocks exógenos de manera más natural. No obstante, presentan el desafío de la calibración: es necesario ajustar los parámetros del modelo para que reproduzca comportamientos plausibles o patrones observados en la realidad, lo cual puede ser complejo dado el gran número de variables intervenientes.

4.3. Evaluación de Políticas Públicas con Econometría

La econometría aplicada ha sido tradicionalmente la herramienta clave para evaluar el impacto de políticas públicas en el mercado inmobiliario. Métodos econométricos rigurosos, como la regresión discontinua o las diferencias en diferencias, permiten identificar relaciones causales entre las intervenciones políticas y las variables del mercado (precios, volumen de transacciones, construcción de vivienda nueva, etc.) bajo ciertos supuestos estadísticos. En esencia, estos métodos comparan la evolución de un indicador en una población afectada por la política vs. una población no afectada (grupo de control), aislando el efecto atribuible a la intervención.

Por ejemplo, se ha analizado cómo una política pública específica –como la regulación de los alquileres turísticos– impacta en la oferta de vivienda disponible para alquiler residencial y en los niveles de renta, comparando zonas o períodos con y sin dicha normativa [13, 10]. Del mismo modo, programas de subvención al alquiler o de vivienda social se han evaluado cuantificando cambios en la asequibilidad entre grupos beneficiarios y no beneficiarios [5].

Estos métodos proporcionan evidencia sólida cuando se cumplen sus supuestos, y suelen ser el estándar para informar decisiones basadas en datos históricos. Sin embargo, tienen limitaciones al extrapolarse a situaciones nuevas: por ejemplo, no pueden predecir fácilmente el efecto de una combinación inédita de políticas, o cambios estructurales en el mercado, ya que dependen de datos pasados. Además, cada análisis econométrico típicamente se enfoca en una política aislada a la vez, dificultando la visión sistémica.

En este proyecto se toma en cuenta la evidencia proveniente de estudios econométricos previos (por ejemplo, elasticidades de oferta y demanda estimadas, o impactos medidos de ciertas regulaciones) para informar el diseño y la calibración del modelo de simulación. De este modo, se combinan las fortalezas de ambos enfoques: la validez causal de los estudios econométricos y la flexibilidad exploratoria de las simulaciones por agentes.

4.4. Estudios Académicos Relevantes y Análisis Geoespacial

Numerosos estudios académicos recientes han aplicado técnicas avanzadas de [Machine Learning \(ML\)](#) y análisis geoespacial al sector inmobiliario, ofreciendo hallazgos útiles para nuestro propósito. Por ejemplo, [Yrigoy \(2017\)](#) [12] analiza los efectos de plataformas como Airbnb en el mercado residencial de destinos turísticos —centrándose en Menorca y comparando sus resultados con los de Barcelona—, encontrando una correlación significativa con fenómenos de gentrificación y la reducción de la oferta de alquiler de larga estancia. Sus resultados apoyan la idea de que, sin una regulación adecuada, la expansión de los alquileres turísticos puede ejercer una presión alcista importante sobre los precios de alquiler tradicionales, desplazando a la población local.

Por otra parte, estudios como **Kortas et al. (2022)** [7] han empleado técnicas de análisis geoespacial para identificar patrones de precios y dinámicas de desplazamiento urbano. Mediante el procesamiento de datos georreferenciados (por ejemplo, utilizando sistemas de información geográfica, GIS), estos trabajos mapean la evolución de los precios inmobiliarios a nivel de barrio o distrito, detectando “zonas calientes” de encarecimiento y anticipando posibles burbujas locales. Tales análisis permiten visualizar cómo ciertas políticas (como limitar la concesión de licencias turísticas en zonas saturadas) podrían enfriar áreas de sobre-demanda y distribuir de forma más uniforme las presiones del mercado.

En la literatura también destacan trabajos como el de Furtado et al. (2018, 2021) [5] –núcleo de la red bibliográfica de la Figura 4: – que introdujeron el modelo PolicySpace original y posteriores extensiones (PolicySpace2). Estas investigaciones integran mercados de bienes, laborales, financieros e inmobiliarios en un solo marco de simulación, algo que la mayoría de estudios previos abordaban por separado. Sus hallazgos subrayan la importancia de considerar las interdependencias: por ejemplo, una política de aumento del salario mínimo puede tener efectos indirectos en el mercado de vivienda (a través de la renta disponible de los hogares y su capacidad de pago), lo que ilustra la necesidad de enfoques integrales como el de este proyecto.

En resumen, el estado del arte proporciona tres lecciones clave: (1) las técnicas de IA y Big Data permiten análisis detallados pero necesitan complementarse con enfoques causales; (2) las simulaciones ABM son prometedoras para explorar escenarios complejos, especialmente inspiradas en modelos existentes validados en la literatura; (3) las evidencias empíricas y estudios de caso documentados sirven de guía y referencia para asegurar que el modelo desarrollado refleje comportamientos realistas y aborda problemas socialmente relevantes. El presente trabajo se apoya en todos estos pilares para proponer una solución innovadora al análisis del sector inmobiliario.

5. Solución Propuesta y Metodología

Este proyecto propone una solución integral para analizar el mercado inmobiliario y evaluar el impacto de las políticas públicas mediante simulaciones basadas en agentes. A continuación, se detallan los elementos fundamentales de la metodología, incluyendo las fuentes de datos, la analítica, la algorítmica, la arquitectura, la visualización y los aspectos éticos y legales.

5.1. Solución Propuesta

El modelo PolicySpace2 (PS2) se adopta como base para este proyecto. PS2 es un modelo computacional basado en agentes (**ABM**) diseñado para simular dinámicas del mercado inmobiliario y el impacto de diversas políticas públicas. La solución propuesta integra mercados de bienes y servicios, trabajo, crédito e inmobiliario, permitiendo observar las interacciones entre agentes como hogares, empresas, bancos y municipios. Además, se comparan políticas como vales de alquiler, ayuda monetaria y distribución de viviendas, evaluando su impacto en términos de producción, consumo y desigualdad.

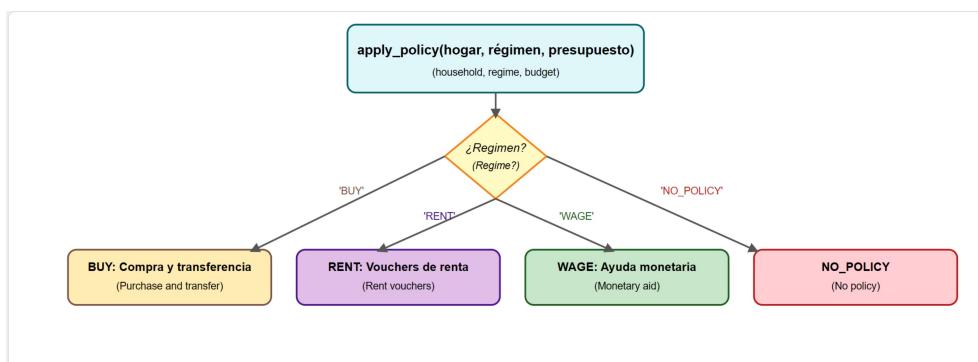


Figura 2: Diagrama de flujo de la aplicación de políticas en el modelo PolicySpace2, mostrando la selección de régimen (Compra, Renta, Ayuda Monetaria, Sin Política). Elaboración del autor.

Para comprender mejor la configuración espacial de los agentes y las entidades dentro del modelo PolicySpace2, se presenta la **Figura 3**, adaptada de la FIGURE 1 del trabajo de Furtado (2022) [5]. Esta figura ilustra cómo se organizan espacialmente los componentes del modelo:



Figura 3: Configuración espacial de los agentes en PolicySpace2 (Adaptado de Furtado, 2022). Elaboración del autor.

La Figura 3 muestra que:

- **Municipios y Regiones (APs):** El modelo se estructura jerárquicamente. Los *Municipios* son las unidades administrativas más grandes, que a su vez se componen de *Regiones* o Áreas de Ponderación (APs). Estas APs representan divisiones intraurbanas, como barrios o distritos, donde se localizan los demás agentes.
- **Familias e Individuos (Agentes):** Los *Agentes* (individuos, trabajadores) se agrupan en *Familias*. Cada familia está asignada a una vivienda específica y, por lo tanto, localizada dentro de una AP particular. Las familias toman decisiones colectivas sobre consumo, ahorro y participación en los mercados.
- **Empresas y Propiedades:** Tanto las *Empresas* (que operan en los mercados de bienes, servicios y laboral) como las *Propiedades* (el stock de viviendas) están también espacialmente asignadas a las APs. Las familias están siempre vinculadas a una dirección, ya sea como propietarias o inquilinas.
- **Banco:** Existe una entidad bancaria única que opera a nivel agregado para todo el sistema simulado, sin una localización espacial específica. Es responsable de la financiación inmobiliaria y de la gestión de los depósitos y ahorros de las familias.

Esta configuración espacial es fundamental en PolicySpace2, ya que las interacciones entre agentes (por ejemplo, el acceso al empleo o a los mercados de bienes) están mediadas por la localización y las distancias, lo que permite analizar dinámicas territoriales complejas.

5.2. Fuentes de Datos

El modelo utiliza datos oficiales y adicionales adaptados a las regiones metropolitanas de estudio en España, incluyendo:

■ Datos demográficos:

- Ubicación y características de individuos y hogares (por ejemplo, tamaño del hogar, estructura familiar).
- Distribución por edad y género.
- Datos de movilidad y migración interna entre regiones.

■ Datos económicos:

- Ubicación de empresas y su actividad económica.
- Estimaciones de población activa y tasa de empleo.
- Índice de desarrollo humano (IDH) y datos sobre desigualdad (e.g., índice de Gini).
- Tasas de interés hipotecarias y datos sobre acceso al crédito.
- Información sobre ingresos medios y dispersión del ingreso por regiones.

■ Datos geográficos:

- Archivos espaciales de municipios y regiones intraurbanas (shapefiles).
- Información geoespacial sobre áreas urbanas, suburbanas y rurales.
- Mapas de uso del suelo y zonas clasificadas para alquiler turístico (No incluidas en las simulaciones).

■ Procesos sociales:

- Mortalidad, fertilidad y patrones de formación de hogares.
- Edad promedio de matrimonio y cambios en la estructura familiar.
- Efectos del envejecimiento demográfico en la demanda de viviendas.

■ **Datos sobre el mercado inmobiliario:**

- Precios de compra y alquiler de viviendas.
- Información sobre el tamaño, la calidad y el tiempo en el mercado de las viviendas.
- Datos históricos sobre la evolución de los precios inmobiliarios.
- Datos sobre pisos turísticos:
 - Ubicación, precios y ocupación promedio.
 - Relación entre alquiler turístico y disponibilidad de viviendas para uso residencial.

■ **Datos fiscales y de políticas públicas:**

- Impuestos municipales sobre propiedades y transferencias del gobierno central.
- Datos sobre subvenciones a la vivienda, regulación del alquiler y ayudas al acceso a viviendas.
- Políticas de limitación o promoción del alquiler turístico.

Es importante destacar que, aunque no todos los datos están disponibles de manera directa, como el precio real de venta de cada vivienda, el modelo puede generar distribuciones de precios y comportamientos basados en datos oficiales y simulaciones. Los precios reales suelen provenir de fuentes privadas o de acceso restringido, como los registros del Colegio de Notarios, agencias urbanísticas, o bases de datos de empresas como TINSA, lo que implica un alto coste o dificultades en su obtención. No obstante, la integración de estas fuentes, en caso de ser accesibles, permitiría una evaluación más precisa del impacto de políticas públicas y dinámicas del mercado inmobiliario en España. En este proyecto se emplearán bases de datos extraídas de Insideairbnb e Idealista con el objetivo de estudio de variables y componentes principales respecto al precio de la vivienda. El objetivo será implementar en un futuro un modelo predictivo preciso en el flujo de precios del modelo según la ubicación de estudio a simular.

5.3. Analítica

El enfoque analítico incluye:

- Modelización de decisiones de agentes, considerando reglas específicas para hogares, empresas y municipios.
- Análisis de múltiples indicadores macroeconómicos como PIB, inflación, desempleo e índice de Gini.
- Calibración de parámetros para garantizar que los resultados se mantengan dentro de valores realistas.
- Análisis de sensibilidad para evaluar la robustez del modelo ante variaciones de parámetros clave.

5.4. Algorítmica

A continuación, se presenta un diagrama que ilustra los procesos secuenciales y las interrelaciones entre los principales agentes y componentes del modelo PolicySpace2. Este diagrama ofrece una visión general de los flujos de decisión e información que se describen con más detalle en esta sección.

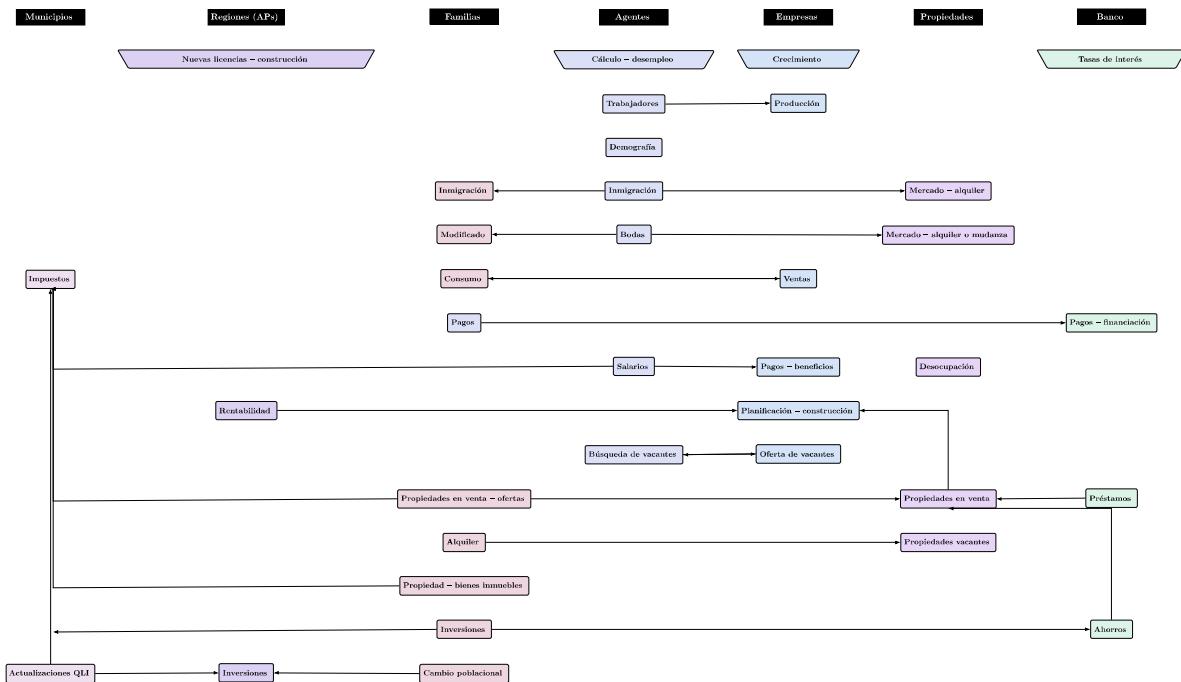


Figura 4: Diagrama de flujo de PolicySpace2: procesos secuenciales e interrelaciones entre agentes (versión en español). Elaboración del autor.

El funcionamiento algorítmico de PolicySpace2 se articula en torno a una secuencia de procesos mensuales que simulan las interacciones de los agentes en diversos mercados y contextos. Una vez inicializados los agentes (familias, empresas, propiedades) y sus atributos, la simulación avanza mes a mes, ejecutando las siguientes etapas en el orden especificado (Furtado, 2022, cap. 3, sec. 4.2) [5]:

1. Actualizaciones Iniciales del Mes:

- Se leen las tasas de interés mensuales y las tasas hipotecarias a partir de datos exógenos.
- Cada región (AP) genera nuevas licencias para la construcción civil (parámetro exógeno).
- Se incorporan nuevas empresas al modelo, manteniendo una proporción basada en datos empíricos y localizándolas probabilísticamente en las APs más dinámicas.

2. Producción y Procesos Demográficos:

- Las empresas actualizan su producción en función del número de empleados y sus cualificaciones (detalles en futuras subsecciones o apéndices).
- Se ejecutan los procesos demográficos (anualmente): envejecimiento, mortalidad y fertilidad (básados en datos oficiales). Los nuevos nacimientos se incorporan a las familias.
- Se procesa la inmigración: se calcula el número anual de migrantes y se distribuye mensualmente. Se generan nuevos agentes y familias, que deben encontrar vivienda en el mercado de alquiler para incorporarse al modelo.
- Se procesan los matrimonios: una proporción de agentes participa, se forman parejas aleatoriamente y, si consiguen alquilar una vivienda, forman nuevas familias o fusionan existentes.

3. Mercado de Bienes y Servicios:

- Las familias toman decisiones de consumo, seleccionando empresas (según proximidad o precio) y determinando la cantidad a gastar en función de su renta permanente.
- El banco cobra las cuotas de las hipotecas.

4. Decisiones de las Empresas y Mercado de Construcción:

- Las empresas evalúan ingresos, pagan impuestos, calculan beneficios/pérdidas, pagan salarios y deciden si actualizan precios.
- Las empresas constructoras planifican nuevas propiedades (considerando rentabilidad y vacancia) y verifican la finalización de construcciones en curso, añadiéndolas a su cartera de ventas.

5. Mercado Laboral:

- Se genera una lista de candidatos (desempleados en edad de trabajar).
- Las empresas (con una frecuencia determinada por un parámetro) deciden si participan en el mercado laboral, pudiendo despedir o abrir nuevas vacantes.
- Se realiza un emparejamiento (matching) entre candidatos y empresas, considerando criterios de cualificación y/o proximidad, costes de transporte y salarios ofrecidos.

6. Mercado Inmobiliario:

- Se actualizan los precios de todas las propiedades. Las propiedades vacantes se listan para venta o alquiler.
- Un porcentaje de familias (parámetro exógeno) más las nuevas familias (por matrimonio o inmigración) entran al mercado.
- Primero opera el mercado de alquiler: las familias buscan según su renta permanente. Si no encuentran, pueden negociar un descuento.
- Luego opera el mercado de compraventa: las familias, ordenadas por poder adquisitivo (incluyendo crédito potencial), buscan propiedades. El precio final se negocia entre la valoración hedónica del comprador y la estimación del vendedor sobre los ahorros del comprador, con posibilidad de descuentos según la vacancia del mercado.
- Las familias realizan inversiones si procede.

7. Gestión Municipal y Cierre del Mes:

- Los municipios invierten los impuestos recaudados en mejoras públicas, lo que afecta al Índice de Calidad de Vida (QLI) local.
- Se recopilan y guardan las estadísticas mensuales del modelo.

Esta secuencia de operaciones, repetida mensualmente, permite la emergencia de dinámicas complejas a partir de reglas de comportamiento microeconómico y interacciones en múltiples mercados interconectados. Los detalles específicos de los submodelos de decisión de los agentes, como la racionalidad de las empresas, el funcionamiento de los mercados de bienes, laboral e inmobiliario, los procesos demográficos y el sistema bancario, se describen con mayor profundidad en el trabajo original de Furtado (2022, cap. 3, sec. 7) [5] y podrían detallarse en futuras subsecciones de este documento si fuera necesario.

5.5. Arquitectura de la Solución

La implementación de un modelo complejo como PolicySpace2, adaptado al contexto español, requiere una arquitectura de software bien definida que gestione la ingesta de datos, el núcleo de simulación basado en agentes y la generación de resultados. Siguiendo los principios de modularidad y transparencia inherentes a muchos modelos ABM (Furtado, 2022) [5], la arquitectura de esta adaptación se estructura en los siguientes componentes principales, como se ilustra conceptualmente en la Figura 5:

La entrada de datos al sistema proviene del módulo ETL que se comentará en la Figura 5, que vuelca la información en una base de datos unificada. Esta base de datos (implementada en PostgreSQL con extensión PostGIS para datos geoespaciales, o alternativamente en archivos CSV/JSON preprocesados para entornos sin SGBD) constituye el “mundo inicial” del modelo: de allí se extraen la lista de municipios con sus atributos, la población inicial de hogares e individuos (sintetizada a partir del censo), el parque de viviendas existente, etc. Antes de iniciar una simulación, se pasa por una etapa de configuración donde el usuario o investigador puede elegir parámetros globales (por ejemplo, simular 100 municipios en lugar de todos, activar o no ciertas políticas, establecer la semilla aleatoria, etc.). Estos parámetros configuran el modelo a través de un archivo de parámetros o una interfaz.

El núcleo de simulación es el corazón de la arquitectura. Está desarrollado en Python siguiendo una estructura modular: existe una clase principal `Simulation` que coordina la ejecución, y clases para agentes principales (`Household`, `Firm`, `Bank`, `Municipality`, etc.) definidas en módulos separados. Asimismo, hay submódulos para procesos específicos, por ejemplo `housing_market.py` maneja las interacciones de oferta-demanda de vivienda, `labor_market.py` el emparejamiento de empleo, `finance.py` las transacciones de crédito, etc. Esta organización facilita la legibilidad y posibilidad de reemplazar componentes (por ejemplo, podría intercambiarse la implementación del mercado inmobiliario por otra más detallada sin afectar el resto). Durante la ejecución, la clase `Simulation` carga los datos iniciales en las estructuras de los agentes, luego entra en un bucle de tiempo donde llama secuencialmente a métodos de actualización en cada submódulo (siguiendo el flujo descrito en la sección sobre Algoritmia).

Un aspecto importante es cómo se representa el entorno espacial: aprovechando los polígonos de municipios cargados, se utiliza la librería `Shapely` para cálculos geométricos (como determinar si una empresa u hogar cae dentro de cierto municipio, o generar puntos aleatorios dentro de un polígono optimizado como se explicó). Esto confiere al modelo una dimensión geográfica explícita que es parte de su arquitectura de datos: básicamente, cada agente tiene una coordenada (x,y) o un polígono asociado, y muchas decisiones (como migración o ubicación de empresas) consideran distancias o pertenencia a zonas definidas.

Otro componente de la arquitectura es el subsistema de políticas: para flexibilidad, se implementó una especie de framework de políticas donde cada política pública posible se codificó como una clase o función que altera ciertos aspectos del modelo. Por ejemplo, una política de “SubsidioAlquilerJoven” es una clase con atributos (monto de ayuda, criterios de elegibilidad) y un método `apply(hogar)` que transfiere el dinero y marca el hogar como beneficiario (afectando su presupuesto). La simulación mantiene una lista de políticas activas y en cada ciclo las aplica según corresponda. Esto hace muy sencilla la prueba de diferentes combinaciones de políticas sin cambiar el código base: solo cambiando la lista de políticas activas en la configuración, el modelo las tendrá en cuenta.

En cuanto a la salida de resultados, el modelo genera gran cantidad de datos a lo largo del tiempo simulado. La arquitectura contempla dos vías principales para gestionar estos resultados:

1. Almacenamiento en base de datos: Al final de cada periodo simulado, se puede guardar el estado relevante en tablas (por ejemplo, distribución de precios, número de transacciones, indicadores agregados por municipio). Esto permite luego consultas directas o conectar herramientas de BI para análisis.
2. Exportación a ficheros e interfaz visual: Alternativamente, se implementó la exportación de resultados a formatos CSV/JSON que alimentan un infograma interactivo desarrollado con idea futura de integrarlo en el programa Streamlit desarrollado. Este infograma se conecta directamente a la simulación: al finalizar una corrida, se lanza o actualiza la aplicación de visualización con los nuevos datos. La arquitectura cliente-servidor aquí es sencilla: la simulación escribe archivos de resultado y app (que corre en Python también) alimenta la web de infograma en la cual se muestran los gráficos.

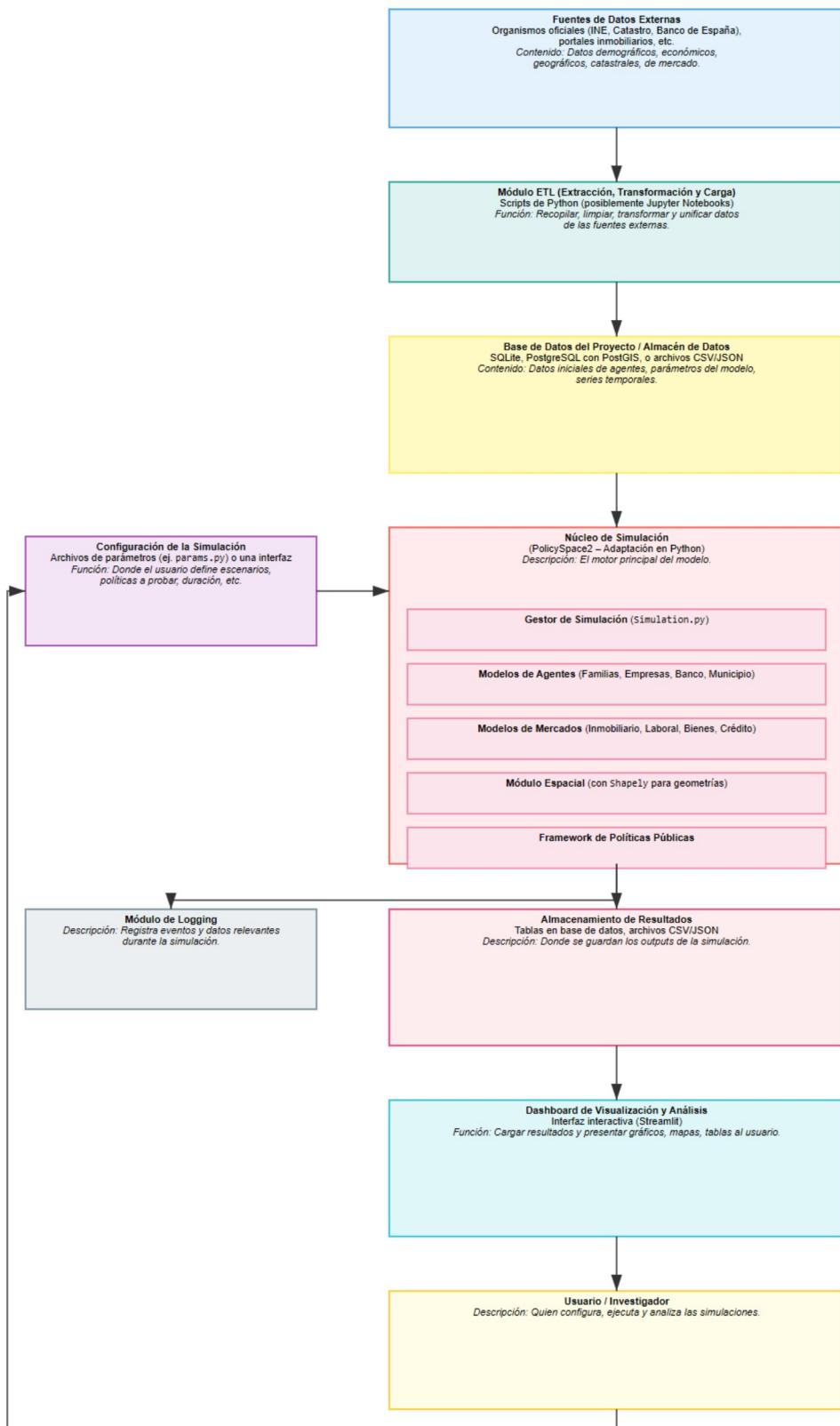


Figura 5: Esquema de los agentes y entidades del modelo y su integración en la arquitectura general. Cada región (municipio) contiene agentes Familias (hogares, compuestos por individuos) y Empresas, así como un conjunto de Propiedades (viviendas). Un único Banco centralizado interactúa con todos los hogares y empresas. Los módulos de datos (ETL) proveen información a los agentes iniciales, y el módulo de visualización (dashboard) se alimenta de los resultados de la simulación. Elaboración del autor.

Por último, se incluyó un módulo de logging y seguimiento que registra los eventos importantes durante la simulación en un archivo de texto (o consola). Esto es útil para depurar y también para explicar determinadas ejecuciones: por ejemplo, loguea cuándo se ejecuta cierta política, o si ocurre algún suceso excepcional (como que un municipio alcance población cero, o que el banco entre en pérdidas, etc.). Estas bitácoras complementan los resultados numéricos con información narrativa. En suma, la arquitectura de la solución está pensada para ser lo más flexible y escalable posible, separando componentes y posibilitando mejorar cada pieza aisladamente. Esto ha permitido, por ejemplo, integrar sin mucha dificultad la gran cantidad de datos exógenos (gracias al módulo ETL) y dotar al proyecto de una interfaz de presentación de resultados sin alterar el núcleo (gracias al diseño modular y la exportación de resultados). La [Figura 5](#) resume visualmente la relación entre los agentes simulados (dominio del modelo) y los componentes técnicos que los soportan.

5.6. Visualización

PS2 genera visualizaciones automáticas que permiten analizar resultados clave, como:

- Gráficos sobre indicadores macroeconómicos.
- Mapas espaciales de la distribución de precios de viviendas y ubicación de agentes.
- Resultados organizados en la carpeta `conf/run`, accesibles para análisis posterior.

5.7. Ética y Legalidad

El proyecto respeta las normativas legales y éticas aplicables al manejo de datos oficiales:

- Cumplimiento del Reglamento General de Protección de Datos RGPD en el manejo de datos sensibles [\[4\]](#).
- Enfoque en la equidad social mediante la comparación de políticas públicas.
- Transparencia y reproducibilidad al mantener los mismos parámetros y procesos para todos los escenarios simulados.

5.8. Resumen Metodológico

La metodología propuesta proporciona una herramienta robusta y replicable para analizar el impacto de políticas públicas en el mercado inmobiliario de España, apoyando la toma de decisiones basada en evidencia empírica y simulaciones dinámicas.

5.9. Estructura de Datos de Referencia y Componentes Iniciales del Proyecto PolicySpace2_Spanish_data

El proyecto `PolicySpace2_Spanish_data` sienta las bases para la adaptación del modelo PolicySpace2 al contexto español. Incluye la definición de equivalencias de datos, una base de datos inicial y un dashboard para la exploración preliminar.

5.9.1. Tablas de Equivalencias y Categorización de Datos

Para mapear los datos del modelo original (contexto brasileño) a fuentes españolas, se definieron tablas de equivalencias. El archivo `equivalencias_datos_espana.csv` lista los documentos de datos originales, su equivalente conceptual en España, la fuente española (principalmente INE, Banco de España, Ministerio de Hacienda) y URLs de acceso. Complementariamente, el documento `categorias_documentos.md` (ubicado en `home/ubuntu/`) categoriza los archivos de datos del proyecto original en secciones como datos geográficos, demográficos (población, fertilidad, mortalidad, matrimonio), económicos (empresas, financiación municipal, indicadores económicos) y educativos. También resume las fuentes principales necesarias para la adaptación

a España. Estos documentos son cruciales para entender la estructura de datos original y guiar la búsqueda y procesamiento de datos españoles.

5.9.2. Base de Datos Inicial y Dashboard

El subproyecto PolicySpace2_Spanish_data incluye una base de datos SQLite inicial (`data_base/datawarehouse.db`) que consolida varios de los datasets procesados por los notebooks ETL. El archivo `data_base/info_tablas.md` describe su estructura, detallando tablas como `tabla_equivalecias`, `cifras_poblacion_municipio`, `df_mortalidad_ccaa`, entre otras, con sus columnas principales y claves. Para la exploración de esta base de datos, se desarrolló un dashboard interactivo utilizando Streamlit (código en `dashboard/app.py`). Esta aplicación se conecta a `datawarehouse.db` y permite:

- Visualizar informes predefinidos basados en los datos procesados (población, mortalidad, IDH, etc.).
- Explorar directamente las tablas de la base de datos.
- Acceder a documentación del proyecto, incluyendo el contenido de `categorias_documentos.md`.

Este dashboard sirve como una herramienta para la validación y análisis preliminar de los datos integrados.

5.10. Proceso de Extracción, Transformación y Carga (ETL) de Datos para España

La adaptación del modelo PolicySpace2 al contexto español requirió un extenso proceso de ETL para recopilar, limpiar y transformar datos de diversas fuentes oficiales españolas (principalmente el INE) y otras fuentes relevantes. A continuación, se resumen los principales procesos ETL realizados, documentados a través del análisis de los notebooks de Jupyter convertidos:

5.10.1. Reducción Progresiva del Conjunto de Datos Municipal

El proceso de ETL no solo se encarga de la ingestión y transformación de datos, sino también de un filtrado progresivo del conjunto de municipios para asegurar la calidad y completitud de la información necesaria para la simulación. La Figura 6 ilustra cómo el número de municipios considerados se reduce en distintas etapas clave del proyecto, partiendo de un conjunto inicial de todos los municipios españoles hasta llegar a un subconjunto con datos consistentes y completos para las variables críticas del modelo, como las liquidaciones presupuestarias de Hacienda y los datos catastrales. Este refinamiento es esencial para la robustez de los análisis posteriores.

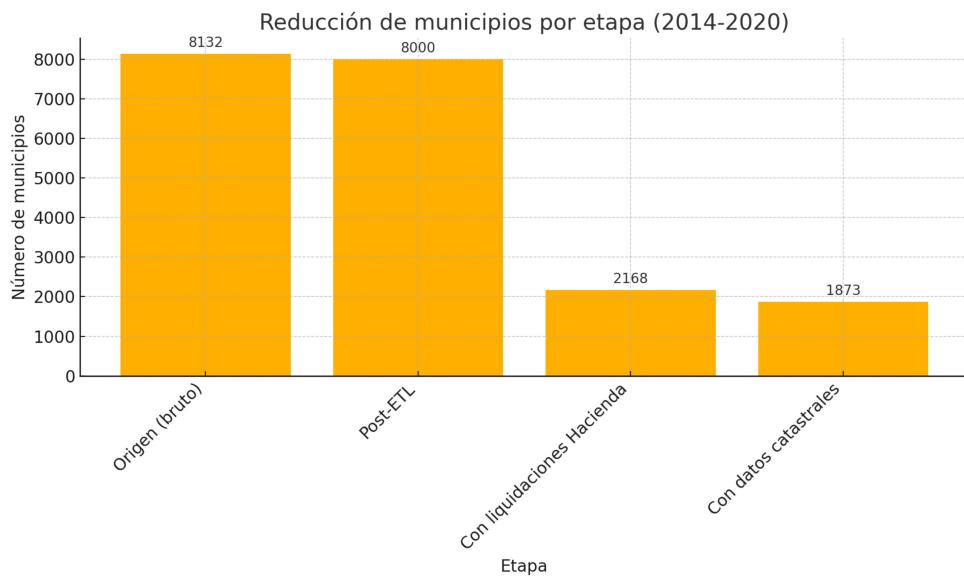


Figura 6: Reducción del número de municipios por etapa del ETL (2014-2020). Elaboración del autor.

5.10.2. Procesamiento de Cifras de Población por Municipio (Versión Completa con Formato Ancho)

El notebook `cifras_poblacion_municipio2_new.ipynb` extiende el procesamiento de la versión `_old`. Además de realizar la limpieza, transformación y segmentación de los datos de población de la tabla 33575 del INE, este notebook añade un paso crucial: la transformación de los datasets anuales por sexo a un formato ancho. En este formato, cada municipio (`cod_mun`) tiene una fila, y las edades (0-100) se convierten en columnas, con los valores de población correspondientes. Este formato replica la estructura del archivo de referencia `pop_men_2000.csv` del proyecto PolicySpace2 original, generando 40 archivos CSV (uno por cada combinación de sexo y año, 2003-2022) listos para ser utilizados por el modelo.

5.10.3. Procesamiento de Datos de Mortalidad por Comunidad Autónoma y Sexo

El notebook `df_mortalidad_ccaa_sexo.ipynb` se encarga de procesar los datos de mortalidad del INE (tabla 27154) para generar tablas de mortalidad por Comunidad Autónoma (CCAA), sexo, edad y año. El proceso incluye filtrar las tasas de mortalidad, limpiar y convertir tipos de datos (edad, tasa de mortalidad), extraer códigos de CCAA, y un manejo detallado de valores NaN y duplicados. Los NaNs, concentrados en edades avanzadas para ciertas CCAA, se imputan utilizando la media nacional (excluyendo Ceuta y Melilla) para la misma edad, sexo y año. Las tasas de mortalidad (originalmente por 1.000 habitantes) se convierten a probabilidades. Finalmente, para cada CCAA y sexo, se generan archivos CSV (38 en total para el periodo 2010-2020) con la edad como índice y los años como columnas, conteniendo las probabilidades de mortalidad, adaptándose así al formato requerido por PolicySpace2.

5.10.4. Estimación de la Proporción de Población Urbana Municipal

El notebook `distribucion_urbana.ipynb` tiene como objetivo estimar la proporción de población urbana para cada municipio de España anualmente (2003-2022). Este proceso integra datos de población total (del notebook `cifras_poblacion_municipio_old.ipynb`), datos de superficie municipal del CNIG, y datos de proporción urbana real del Nomenclátor del INE para 2016 (usados para calibración). Se realiza un complejo emparejamiento de municipios entre fuentes (incluyendo *fuzzy matching* con `rapiddfuzz`), se calcula la densidad poblacional y se aplica un conjunto de reglas heurísticas para asignar una proporción urbana estimada. Estas reglas consideran umbrales de densidad y población total, refinándose con la proporción urbana real de 2016 para evitar sobreestimaciones. El resultado es un archivo CSV

(`distribucion_urbana_municipios_2003_to_2022.csv`) con la proporción urbana estimada para cada municipio en formato ancho (años como columnas), un insumo clave para el modelo PolicySpace2.

5.10.5. Procesamiento del Número de Empresas por Municipio y Actividad Principal

El notebook `empresas_municipio_actividad_principal.ipynb` procesa datos del INE (tabla 4721) sobre el número de empresas por municipio y actividad principal para el periodo 2012-2024. El objetivo es obtener una serie temporal del número total de empresas por municipio. El ETL incluye la descarga, limpieza inicial, filtrado para obtener el "Total" de empresas por actividad, y la extracción de códigos y nombres de municipio. Una parte crucial es la imputación de valores NaN en el número de empresas: se eliminan algunos municipios sin datos, se imputan ceros a municipios con baja población y actividad persistente nula, y los NaNs restantes se imputan con la media del propio municipio a lo largo de los años con datos. El resultado es el archivo `empresas_municipio_actividad_principal.csv`, con el número total de empresas por municipio y año, sin valores NaN.

5.10.6. Generación de Estimaciones de Población Municipal Anual (Formato Ancho)

El notebook `estimativas_pop-v2.ipynb` se dedica a procesar y limpiar datos de población municipal para generar series temporales anuales (1996-2024) por municipio. Partiendo de un archivo consolidado de población (`df_unido.csv`), se filtran los datos para el total de población, se extraen códigos municipales y se pivota el DataFrame para obtener un formato ancho (años como columnas). Un paso importante es la detección y corrección de outliers en las series temporales de población de cada municipio, utilizando el método del Rango Intercuartílico (IQR) y reemplazando outliers con la media de vecinos válidos. Posteriormente, se eliminan municipios con un exceso de valores NaN persistentes y los NaNs restantes se rellenan mediante interpolación lineal y *forward/backward fill*. El producto final es `cifras_poblacion_municipio.csv`, un dataset con estimaciones de población anual por municipio, sin NaNs y en formato ancho, similar al archivo de referencia `estimativas_pop.csv` del proyecto PolicySpace2 original.

5.10.7. Procesamiento de Indicadores de Fecundidad por Provincia y Comunidad Autónoma

El notebook `indicadores_fecundidad_municipio_provincias.ipynb` transforma los datos de indicadores de fecundidad del INE (tabla 29295), que están disponibles por provincia y grupos de edad. El proceso incluye la limpieza de datos, la extracción de códigos de provincia y la obtención de la edad inicial y final de cada grupo. Un paso clave es la interpolación lineal de las tasas de fecundidad (originalmente por 1.000 mujeres) para estimar tasas por edad individual (10-50 años). Estas tasas se estandarizan (nacimientos por mujer) y luego se agregan para obtener tasas promedio por Comunidad Autónoma. Finalmente, los datos se pivotan a un formato ancho (edad como índice, años como columnas) y se exportan en archivos CSV separados para cada provincia y cada CCAA, listos para su uso en el modelo PolicySpace2.

5.10.8. Procesamiento de Datos de Tasas de Interés

El notebook `interest_data_ETL.ipynb` se encarga de obtener, procesar e imputar datos de tasas de interés mensuales para España (enero 2000 - abril 2025). Utiliza datos del BCE (tasa de interés oficial), del Banco de España (tasa hipotecaria) y de Eurostat (inflación HICP). El proceso genera tres archivos CSV:

- `interest_fixed.csv`: Tasas de interés y hipotecarias fijas (valor constante del 5 % anual).
- `interest_nominal.csv`: Tasas de interés base e hipotecarias nominales, descargadas y remuestreadas mensualmente. Los valores NaN se imputan mediante *forward fill*.
- `interest_real.csv`: Tasas de interés base e hipotecarias reales, calculadas restando la tasa de inflación intermensual (derivada del HICP) a las tasas nominales. Los NaNs resultantes también se imputan con *forward fill*.

El notebook incluye análisis de patrones temporales y validaciones para asegurar la coherencia de los datos imputados.

5.10.9. Procesamiento de Datos de Nivel Educativo por Comunidad Autónoma

El notebook `nivel_educativo_comunidades.ipynb` procesa datos trimestrales del INE (tabla 65289) sobre el nivel de formación de la población por Comunidad Autónoma para el periodo 2014-2024. El ETL incluye la limpieza de datos, la extracción de códigos de CCAA, el filtrado para Ambos sexos, y la agregación de los datos trimestrales a una media anual. Los niveles de formación textuales se codifican en 7 categorías numéricas. Finalmente, los datos se pivotan para obtener un formato ancho y se exportan en archivos CSV anuales (ej., `nivel_educativo_comunidades2014.csv`), donde cada archivo contiene el porcentaje promedio de población en cada nivel educativo para cada CCAA.

5.10.10. Procesamiento de Datos de Participación en los Ingresos del Estado (PIE)

El notebook `PIE_fin_procesamiento.ipynb` se enfoca en el procesamiento de datos de la Participación en los Ingresos del Estado (PIE) a nivel municipal para el período 2007-2022. Partiendo de un archivo Excel (`pie_final.xlsx`), se generan códigos municipales estandarizados y se filtran los municipios para asegurar que tengan una serie temporal completa de 16 observaciones anuales. Los códigos municipales se validan contra el diccionario oficial del INE. El resultado principal es el archivo `PIE_total_2007_2022.csv`, que contiene la PIE por municipio y año, además de archivos desglosados por provincia.

5.10.11. Procesamiento de Datos de Tamaño Medio de los Hogares por Comunidad Autónoma

El notebook `tamaño_medio_hogares_ccaa.ipynb` procesa datos trimestrales del INE (tabla 60132) sobre el tamaño medio de los hogares por Comunidad Autónoma para el período 2021-2025. El ETL incluye la limpieza de datos, la transformación de fechas, la unión con códigos de CCAA y la agregación de los datos trimestrales a una media anual. Los datos finales se exportan en archivos CSV anuales (ej., `tamaño_medio_hogares_ccaa_2021.csv`), cada uno conteniendo el tamaño medio del hogar para cada CCAA.

Cuadro 1: Índice de notebooks y recursos principales del proyecto.

Descripción	Datasets principales	Enlace GitHub / Script
Cálculo y visualización de tasas de fertilidad municipales y autonómicas desde 1975.	tasas_fertilidad_provincias, tasas_fertilidad_comunidades	GitHub
Procesado ETL y análisis del padrón municipal. Incluye generación de dataset final para simulación.	pie_final_final.csv	GitHub
Análisis de mortalidad por comunidad y sexo. Comparativa anual 2010–2020.	mortality_SEXO_COMUNIDAD_CODE.csv	GitHub
Equivalencia de códigos y correspondencias entre territorios de Brasil y España para PolicySpace2.	—	GitHub
Obtención y visualización del tamaño medio de hogar por comunidad autónoma.	tamaño_medio_hogares_ccaa_AÑOX	GitHub
ETL y análisis de cifras de población municipal.	cifras_poblacion_municipio.csv	GitHub
Procesamiento de empresas por municipio y rama principal de actividad (2012–2024).	empresas_actividad_principal.csv	GitHub
Cálculo y evolución del IDH municipal.	idhm_2013_2022.csv	GitHub
Análisis y ETL de tipos de interés (fijo, nominal, real) desde 2000.	interest_fixed_imputado.csv, interest_nominal_imputado.csv, interest_real_imputado.csv	GitHub
Cifras de población desglosadas por sexo y año en municipios (2003–2022).	cifras_municipio_hombres_AÑOX.csv, cifras_municipio_mujeres_AÑOX.csv	GitHub
Procesamiento de datos sobre nivel educativo medio por comunidad (2014–2025).	nivel_educativo_comunidades_AÑOX	GitHub

5.10.12. Análisis Exploratorio Visual de Datos de Entrada

Durante el proceso ETL y la preparación de los datos de entrada para el modelo, se realizaron diversos análisis exploratorios para comprender mejor las características, distribuciones y relaciones inherentes a las variables clave. Estas visualizaciones son fundamentales para identificar patrones, posibles outliers, y la calidad general de los datos que alimentarán la simulación. A continuación, se presentan algunas de estas gráficas exploratorias generadas a partir de los datasets procesados.

Distribución del IDHM por Comunidad Autónoma. El Índice de Desarrollo Humano Municipal (IDHM) es un indicador sintético crucial que refleja las condiciones socioeconómicas a nivel local, considerando aspectos de salud, educación e ingresos. La Figura 7 muestra la distribución de este índice para los municipios dentro de cada Comunidad Autónoma (CCAA) en el año 2022. Se puede observar la mediana, los cuartiles y la dispersión general del IDHM, lo que permite apreciar la heterogeneidad del desarrollo humano tanto dentro de cada CCAA como entre las distintas comunidades autónomas de España. Esta variabilidad es un factor importante que el modelo PolicySpace2 busca capturar y analizar en sus simulaciones.

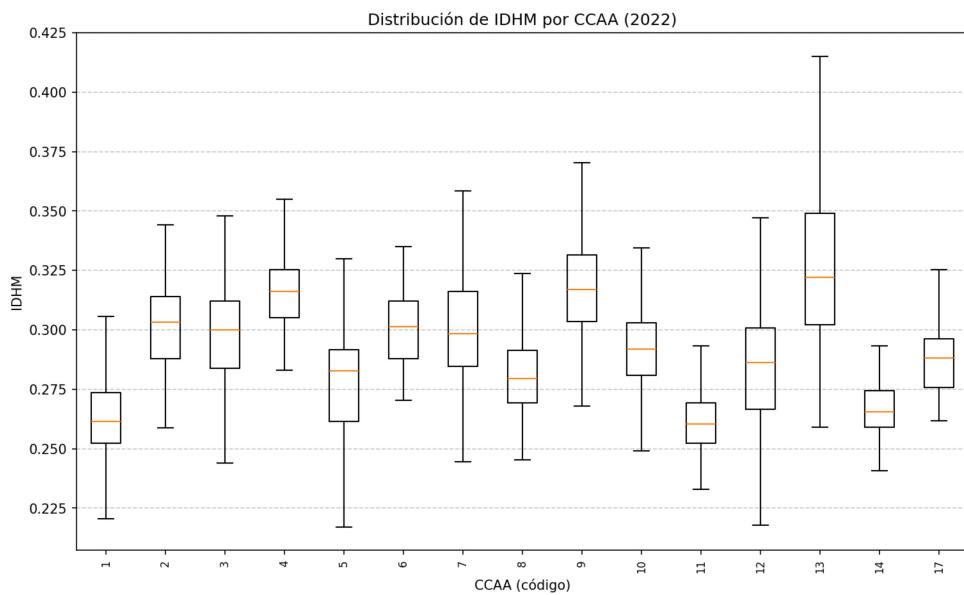


Figura 7: Distribución del IDHM municipal por Comunidad Autónoma (2022). Los códigos numéricos en el eje X representan a las diferentes CCAA según la codificación estándar. Elaboración del autor.

Evolución de la Correlación de Subíndices del IDHM. Para entender la dinámica interna del IDHM, la Figura 8 presenta la evolución de la correlación de cada uno de sus subíndices componentes (ingresos, salud y educación) con el índice general a lo largo del tiempo (aproximadamente 2013-2022). Se observa que el subíndice de ingresos (*I_ingresos*) mantiene una correlación muy alta y estable. En contraste, la correlación del subíndice de educación (*I_educ*) con el IDHM general ha mostrado una tendencia decreciente, lo que podría indicar cambios en la contribución relativa de este componente al desarrollo humano global medido o diferencias en las tasas de mejora entre los componentes. El subíndice de salud (*I_salud*) presenta una correlación moderada con fluctuaciones.

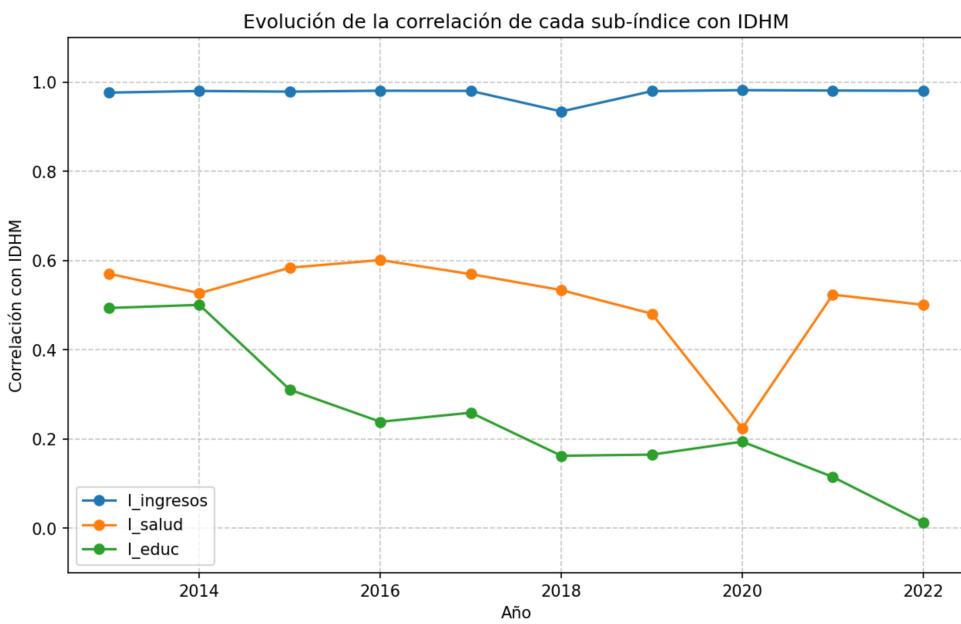


Figura 8: Evolución de la correlación de los subíndices de Ingresos, Salud y Educación con el IDHM general (aprox. 2013-2022). Elaboración del autor.

Distribución General del IDHM Municipal. Complementando el análisis por CCAA y la evolución de sus componentes, la Figura 9 presenta un histograma de la distribución de los valores del IDHM para el conjunto de todos los municipios españoles en 2022. Esta visualización permite apreciar la forma general de la distribución, su tendencia central y dispersión a nivel nacional. Se observa una concentración de municipios en el rango de IDHM entre 0.25 y 0.35, con una forma que se aproxima a una distribución normal aunque con una ligera asimetría positiva, indicando la presencia de algunos municipios con valores de desarrollo humano notablemente más altos.

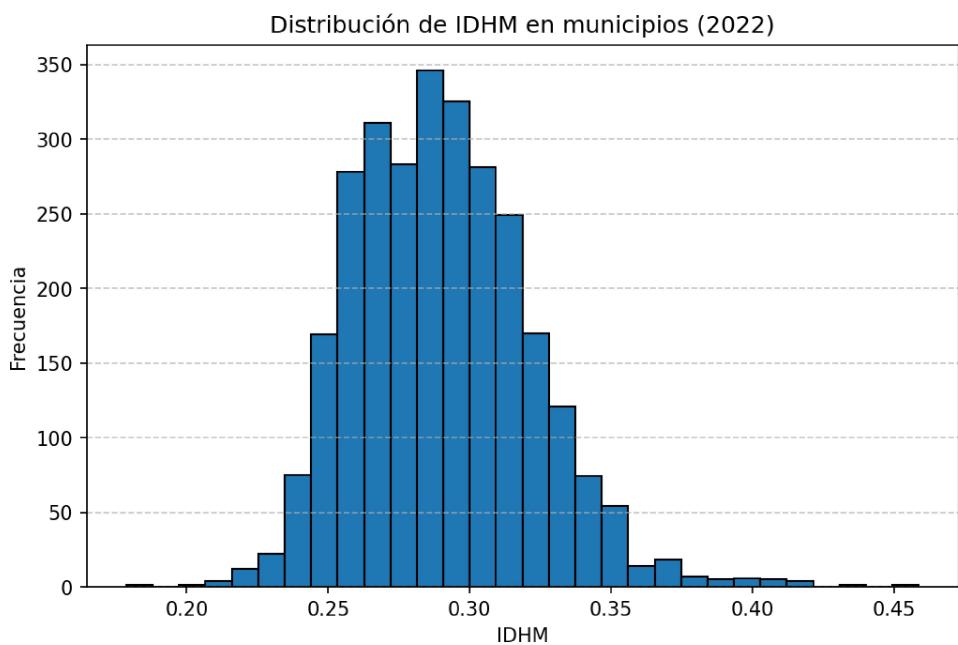


Figura 9: Histograma de la distribución del IDHM en los municipios de España (2022). Elaboración del autor.

Evolución de las Tasas de Interés. Las condiciones financieras, y en particular las tasas de interés, son un factor determinante en el mercado inmobiliario y en las decisiones de los agentes económicos. Las siguientes figuras muestran la evolución de diferentes tasas de interés relevantes para España durante el periodo aproximado 2000-2025, comparando las series originales con las versiones imputadas utilizadas en el modelo.

La Figura 10 muestra la evolución de la tasa de interés real. Se puede apreciar la considerable volatilidad de esta tasa, especialmente después de la crisis financiera de 2008, con periodos donde incluso alcanzó valores negativos, cambiando la tendencia tras el Covid.

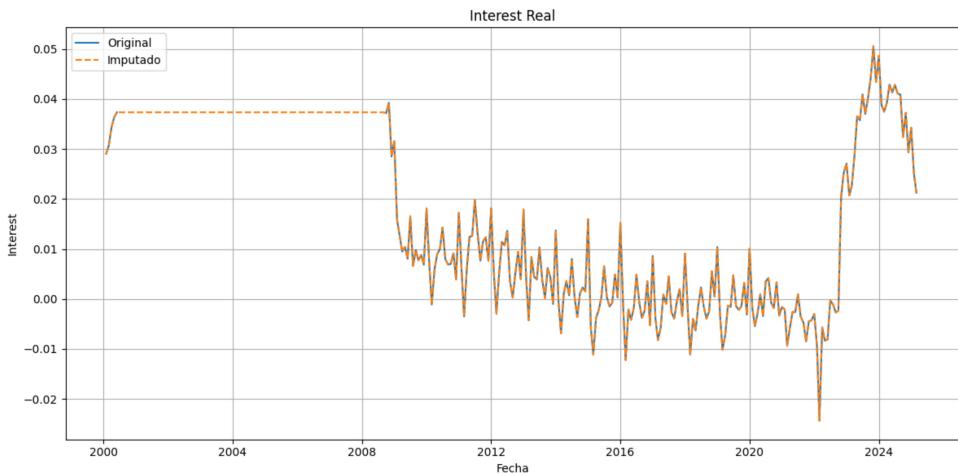


Figura 10: Evolución de la tasa de interés real en España (aprox. 2000-2025), mostrando la serie original y la imputada. Elaboración del autor.

Por su parte, la Figura 11 detalla la evolución de la tasa de interés hipotecaria nominal. Esta tasa también exhibe ciclos claros, con picos en 2000-2001 y 2008-2009, seguidos de un periodo de tasas notablemente bajas entre 2015 y 2022, y un repunte más reciente. La comprensión de estas dinámicas es fundamental para modelar la accesibilidad al crédito hipotecario.



Figura 11: Evolución de la tasa de interés hipotecaria nominal en España (aprox. 2000-2025), original e imputada. Elaboración del autor.

Finalmente, la Figura 12 complementa este análisis mostrando la tasa de interés hipotecaria real. Al igual que la tasa de interés real general, esta también presenta una notable volatilidad y períodos con valores negativos, reflejando el coste efectivo del financiamiento hipotecario una vez descontada la inflación.



Figura 12: Evolución de la tasa de interés hipotecaria real en España (aprox. 2000-2025), original e imputada. Elaboración del autor.

Relación entre IDHM y Tamaño Poblacional. Para explorar cómo el tamaño poblacional de los municipios se relaciona con sus niveles de desarrollo humano, se presentan dos visualizaciones complementarias para el año 2022. Primero, la Figura 13 muestra estimaciones de densidad kernel (KDE) de la distribución del IDHM, agrupando los municipios en cuatro cuartiles según su población (Q1 el más bajo, Q4 el más alto). Las curvas sugieren que, si bien existe una considerable superposición, los municipios en cuartiles de población más altos tienden a mostrar distribuciones de IDHM con modas ligeramente desplazadas hacia valores superiores.

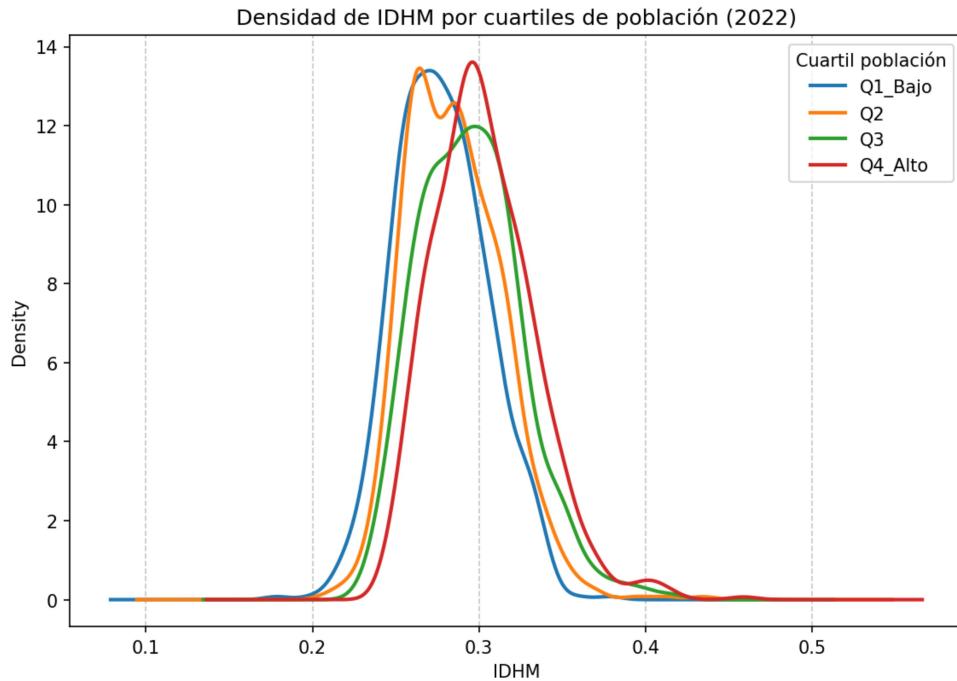


Figura 13: Estimación de Densidad Kernel (KDE) de la distribución del IDHM por cuartiles de población municipal (2022). Elaboración del autor.

Complementando esta visión, la Figura 14 presenta un diagrama de dispersión que relaciona directamente el IDHM de cada municipio con el logaritmo de su población. Esta gráfica confirma la tendencia positiva: a medida que aumenta el tamaño poblacional, el IDHM tiende a ser mayor, aunque con una dispersión que se incrementa en los municipios más grandes, indicando una mayor variabilidad del desarrollo humano en estos casos.

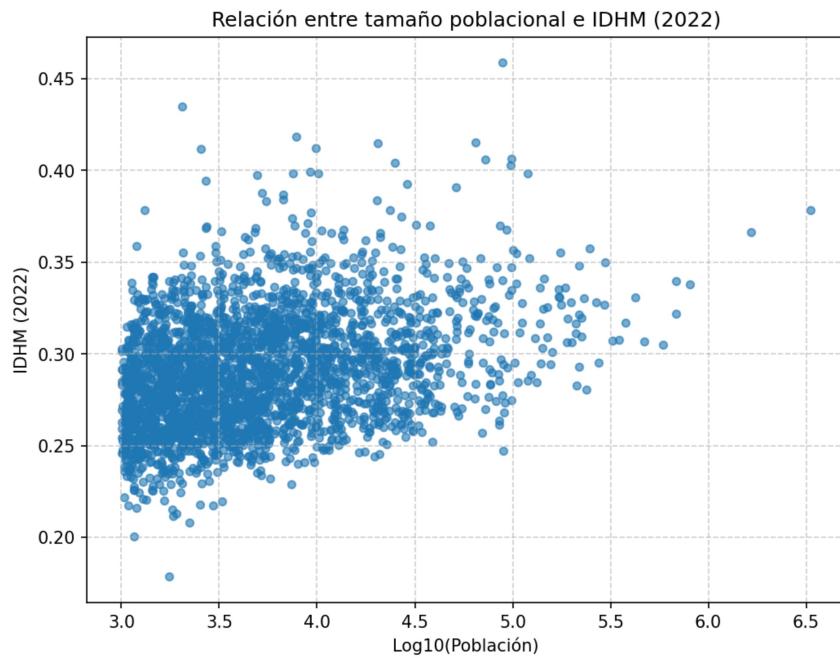


Figura 14: Diagrama de dispersión: Logaritmo de la Población vs. IDHM municipal (2022). Elaboración del autor.

Relación entre Subíndices del IDHM. Para profundizar en la estructura del IDHM, se examinaron las relaciones bivariadas entre sus componentes (educación, salud e ingresos) para el año 2022. La Figura 15 muestra la relación entre el subíndice de educación y el de salud. Aunque no se aprecia una correlación lineal simple y fuerte a primera vista, la dispersión de los puntos (que podrían representar agregados regionales como CCAA) sugiere que existen diferentes combinaciones de niveles de desarrollo en estas dos dimensiones.

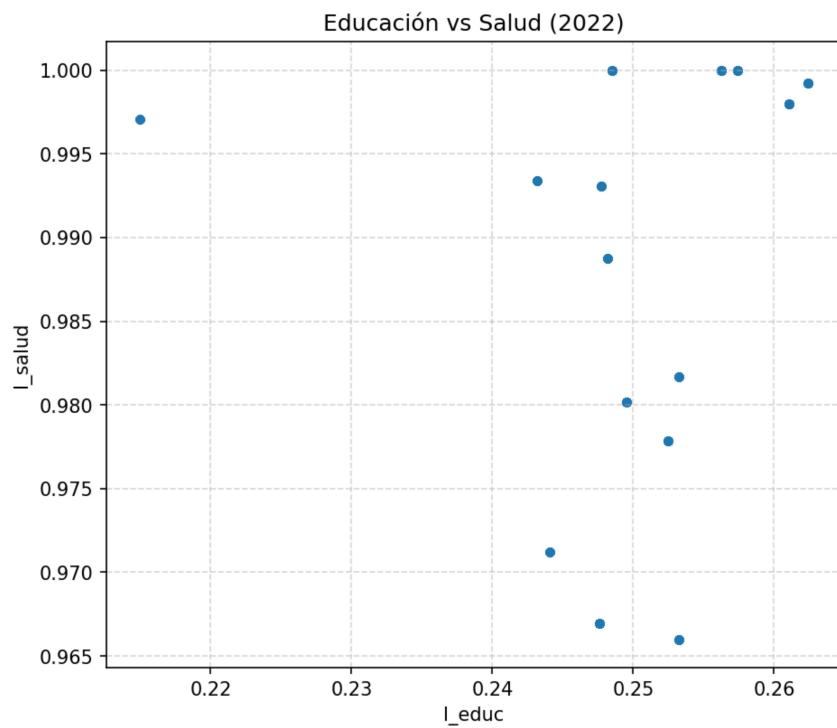


Figura 15: Diagrama de dispersión: Subíndice de Educación vs. Subíndice de Salud del IDHM (2022). Elaboración del autor.

De manera similar, la Figura 16 explora la relación entre el subíndice de ingresos y el de educación. En este caso, se observa una tendencia positiva más clara: a mayores niveles de ingresos, generalmente corresponden mayores niveles en el subíndice de educación, aunque con una dispersión considerable que indica que otros factores también influyen.

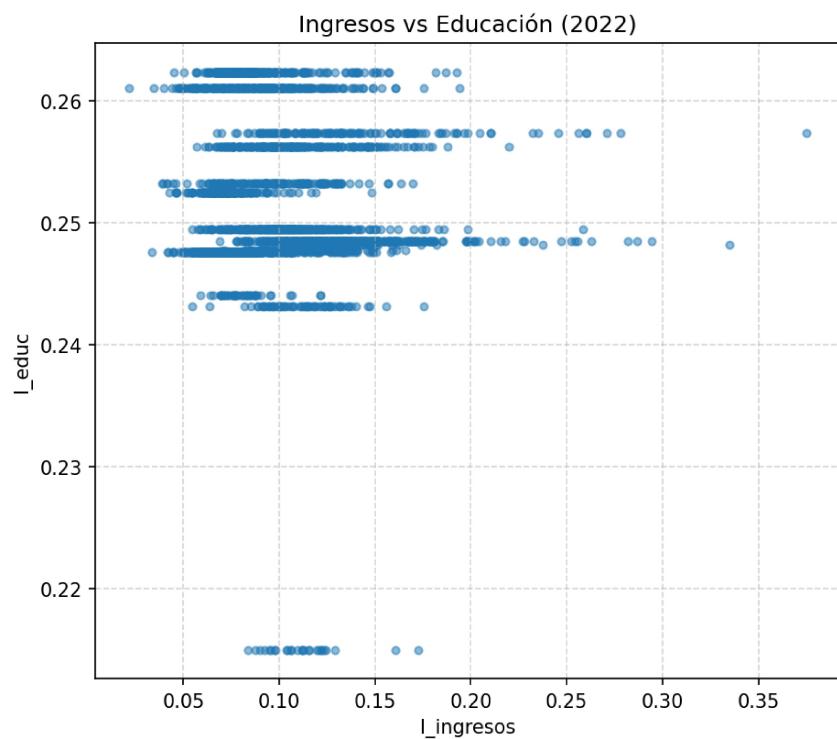


Figura 16: Diagrama de dispersión: Subíndice de Ingresos vs. Subíndice de Educación del IDHM (2022). Elaboración del autor.

Finalmente, la Figura 17 muestra la relación entre los subíndices de ingresos y salud. Aquí también se sugiere una tendencia positiva: mayores ingresos tienden a asociarse con mejores niveles de salud, aunque la relación presenta una dispersión notable y la posible formación de agrupaciones o "bandas" horizontales en los niveles de salud, lo que podría indicar umbrales o la influencia de otros factores no visualizados. Estos diagramas de dispersión ayudan a comprender las interdependencias (o la falta de ellas) entre las dimensiones que componen el IDHM.

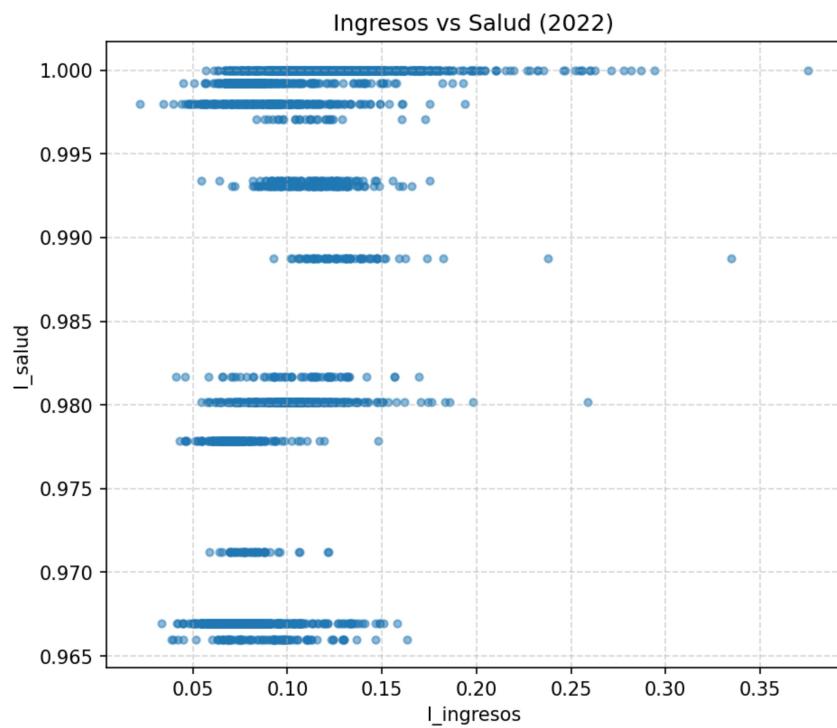


Figura 17: Diagrama de dispersión: Subíndice de Ingresos vs. Subíndice de Salud del IDHM (2022). Elaboración del autor.

Para una visión más integrada, la Figura 18 presenta un diagrama de dispersión multivariado. En este gráfico, la relación entre ingresos y educación se mantiene en los ejes X e Y respectivamente, mientras que el color de los puntos indica el nivel del subíndice de salud y el tamaño de los puntos representa el logaritmo de la población municipal. Esta visualización permite apreciar cómo los municipios con mayores ingresos y niveles educativos tienden también a mostrar mejores indicadores de salud (colores más claros) y a ser más poblados (puntos más grandes), reforzando la idea de una interconexión entre estas dimensiones del desarrollo.

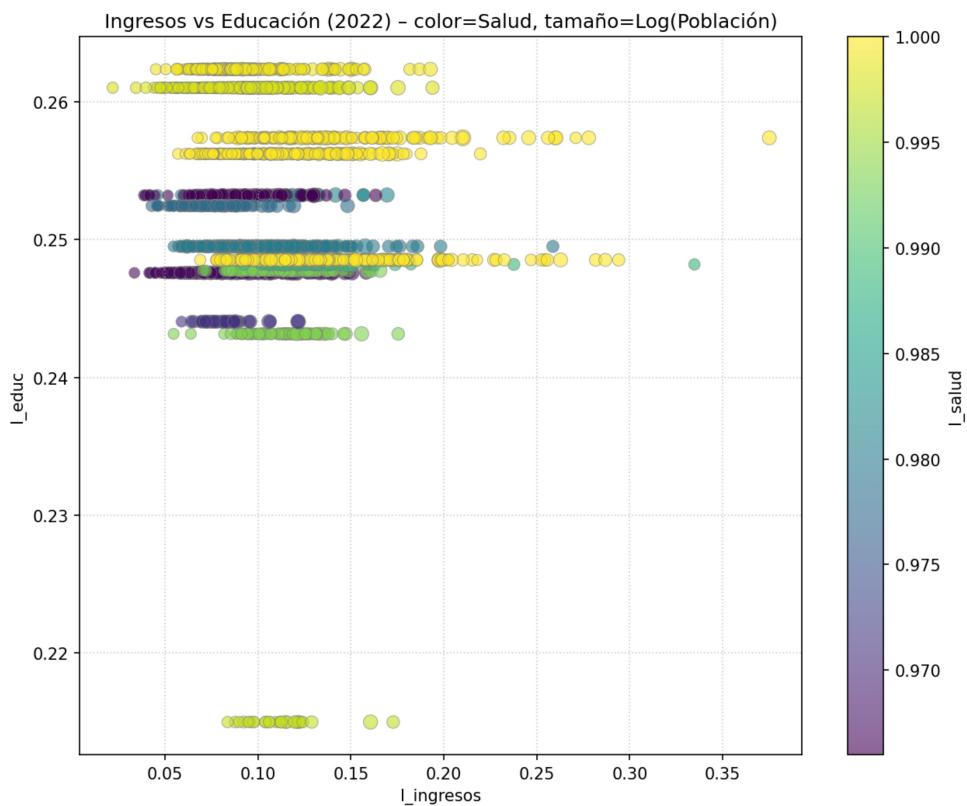


Figura 18: Diagrama de dispersión multivariado: Ingresos vs. Educación, coloreado por Salud y tamaño por Log(Población) (2022). Elaboración del autor.

Evolución Temporal del IDHM Nacional. Además de las relaciones transversales, es importante observar la evolución del IDHM a lo largo del tiempo. La Figura 19 muestra la serie temporal del IDHM promedio nacional para el periodo 2013-2022. Se observa una tendencia general al alza, indicando una mejora progresiva en el desarrollo humano promedio a nivel nacional, con una ligera interrupción o estancamiento alrededor del año 2020, que podría estar asociada a los efectos de la pandemia.

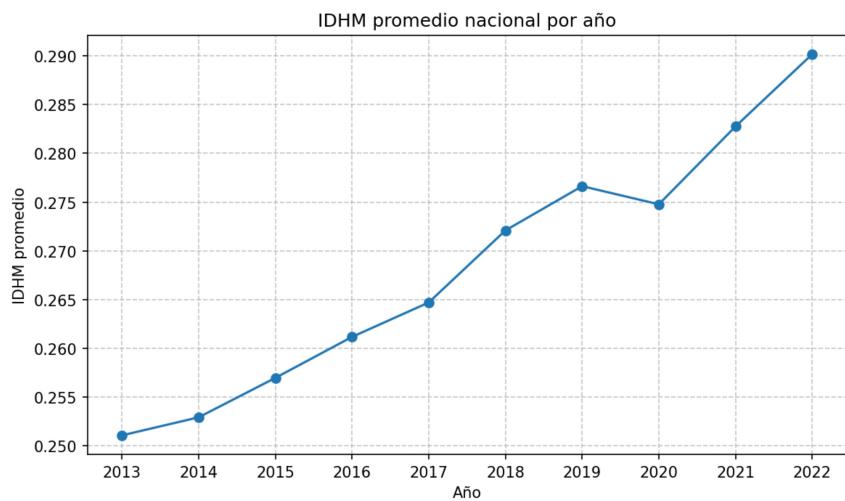


Figura 19: Evolución del IDHM promedio nacional por año (2013-2022). Elaboración del autor.

Para contextualizar esta evolución nacional, la Figura 20 compara la tendencia del IDHM promedio nacional con la de tres de las Comunidades Autónomas (identificadas por sus códigos 13: Madrid, 9: Cataluña, 4: Islas Baleares) que consistentemente muestran un IDHM promedio superior. Esta comparación resalta la heterogeneidad regional en el desarrollo humano y cómo, a pesar de una mejora general, las brechas entre el promedio nacional y las regiones de mayor desarrollo persisten a lo largo del tiempo.

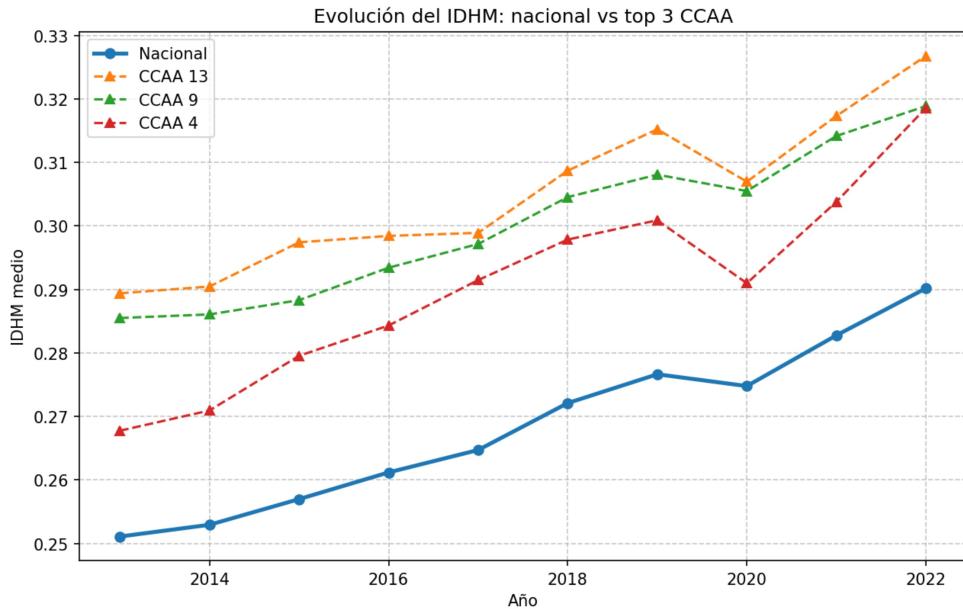


Figura 20: Evolución del IDHM promedio: Nacional vs. Top 3 CCAA (2013-2022). Elaboración del autor.

5.11. Adaptación del Modelo PolicySpace2 y Componentes Adicionales

La adaptación del modelo PolicySpace2 al contexto español implicó varios desarrollos adicionales, incluyendo la mejora del ETL, la integración de nuevas fuentes de datos como el catastro, la creación de modelos predictivos específicos y la generación de análisis de resultados.

5.11.1. Adecuación del Modelo y Base de Datos Mejorada (PolicySpace2_Spain_new_ETL)

El proyecto PolicySpace2_Spain_new_ETL se centró en refinar la adaptación del modelo. Documentos como `adapter_spain_info.md` y `Plan de Adaptación del Módulo de Viviendas (PolicySpace-España).pdf` detallan este proceso. Se abordaron problemas en la generación de agentes y se ajustó la lógica de carga de datos de población y viviendas.

Se desarrolló una base de datos mejorada (`base_datos_mejorada/base_datos_mejorada.db`), cuya estructura se describe en `README_BASE_DATOS.md` y `esquema_propuesto.md`. Esta base de datos consolida los diversos datasets procesados y sirve como fuente central para la simulación y análisis. Incluye tablas dimensionales geográficas y tablas de hechos para población, IDH, PIE, empresas, mortalidad, fecundidad, tipos de interés, nivel educativo, tamaño de hogares y viviendas catastrales. Se generaron esquemas visuales de esta base de datos mediante herramientas como Graphviz y Eralchemy.

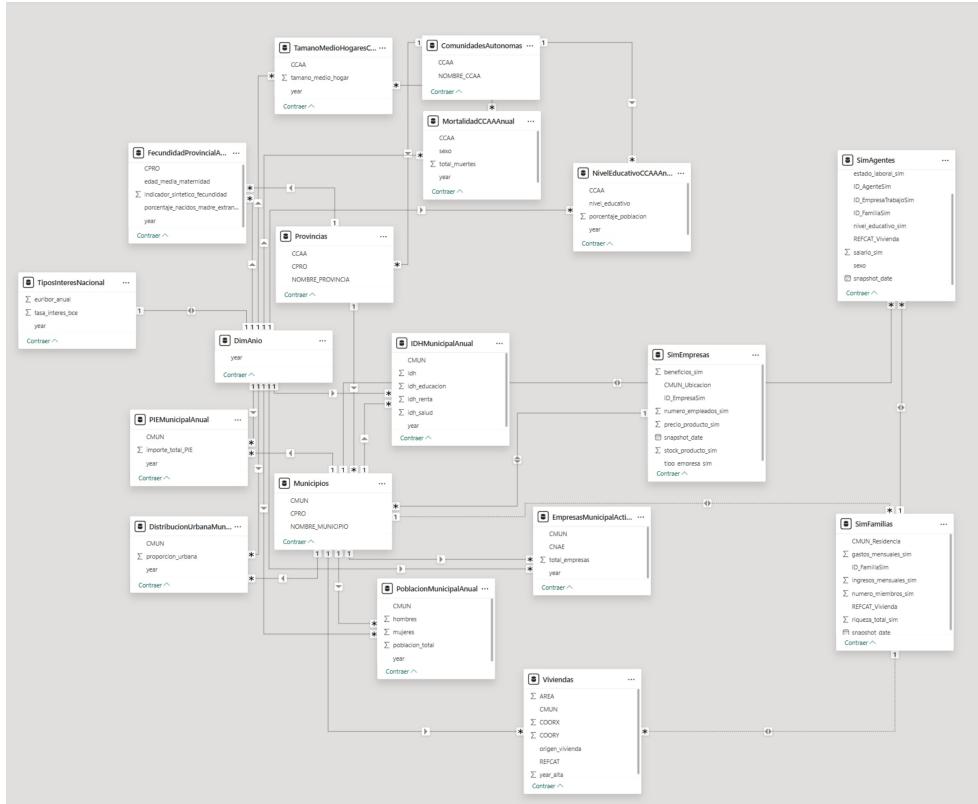


Figura 21: Esquema relacional de la base de datos mejorada del proyecto. Elaboración del autor.

Un componente clave de esta fase fue la creación del dataset `dataset_municipio_cnae_anual_2014_2020.csv`, que integra múltiples variables a nivel municipal para un conjunto de municipios con información completa.

5.11.2. Integración de Datos Catastrales y Generación de Viviendas

La adaptación española de PolicySpace2 incorpora un sistema dual para la generación y ubicación de viviendas, buscando el mayor realismo posible mediante el uso de datos catastrales cuando están disponibles, y recurriendo a una generación sintética espacialmente consciente en su ausencia. Esta lógica se implementa principalmente en el módulo `world/generator.py` del proyecto `PolicySpace2_Spain_new_ETL`.

El modelo decide qué método utilizar basándose en el parámetro de configuración `USE_CADASTRAL_HOUSES` y una lista predefinida de provincias donde los datos catastrales podrían no estar disponibles o no ser utilizables (`PROVINCES_NO_CADASTRO`, que incluye, por ejemplo, Álava, Gipuzkoa, Navarra y Bizkaia). La Figura 22 ilustra conceptualmente estos dos escenarios.

Uso de Datos Poligonales y Generación de Viviendas Sintéticas

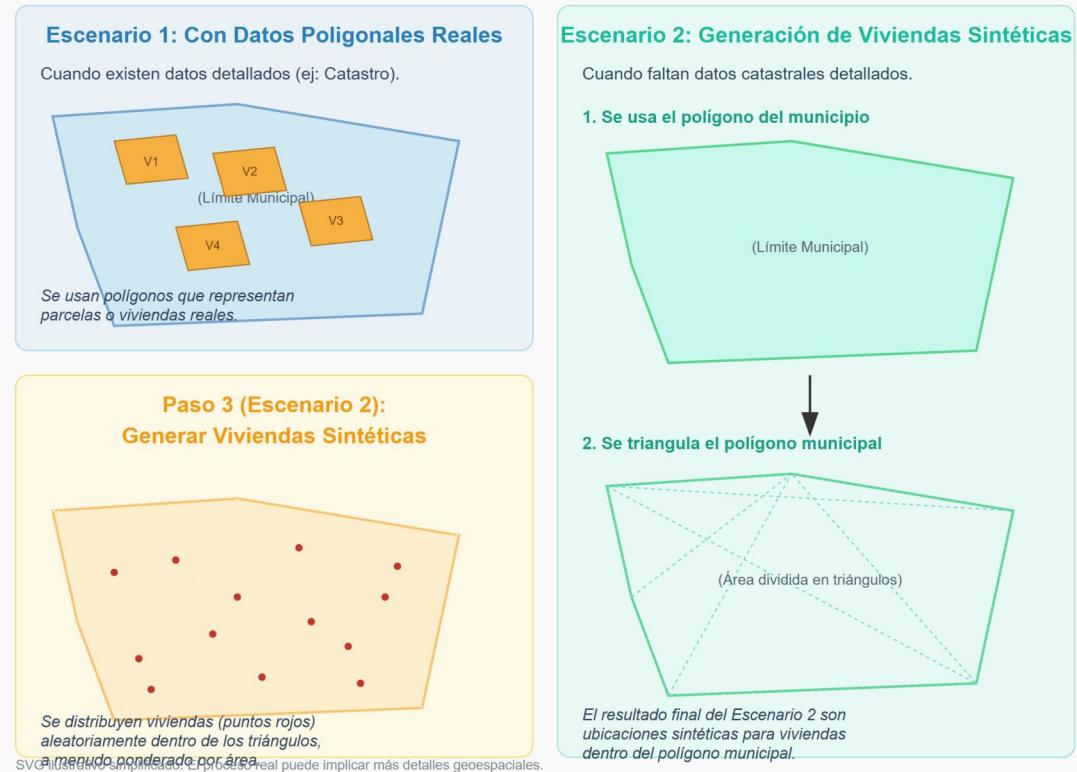


Figura 22: Comparación conceptual entre la generación de viviendas con datos catastrales reales (Escenario 1) y la generación sintética (Escenario 2). Elaboración del autor.

Escenario 1: Uso de Datos Catastrales Reales. El Escenario 1 se activa cuando se dispone de datos catastrales detallados. La Figura 23 muestra la cobertura geográfica de los datos de viviendas catastrales utilizados en este proyecto, que sirven como base para la localización precisa de las propiedades.

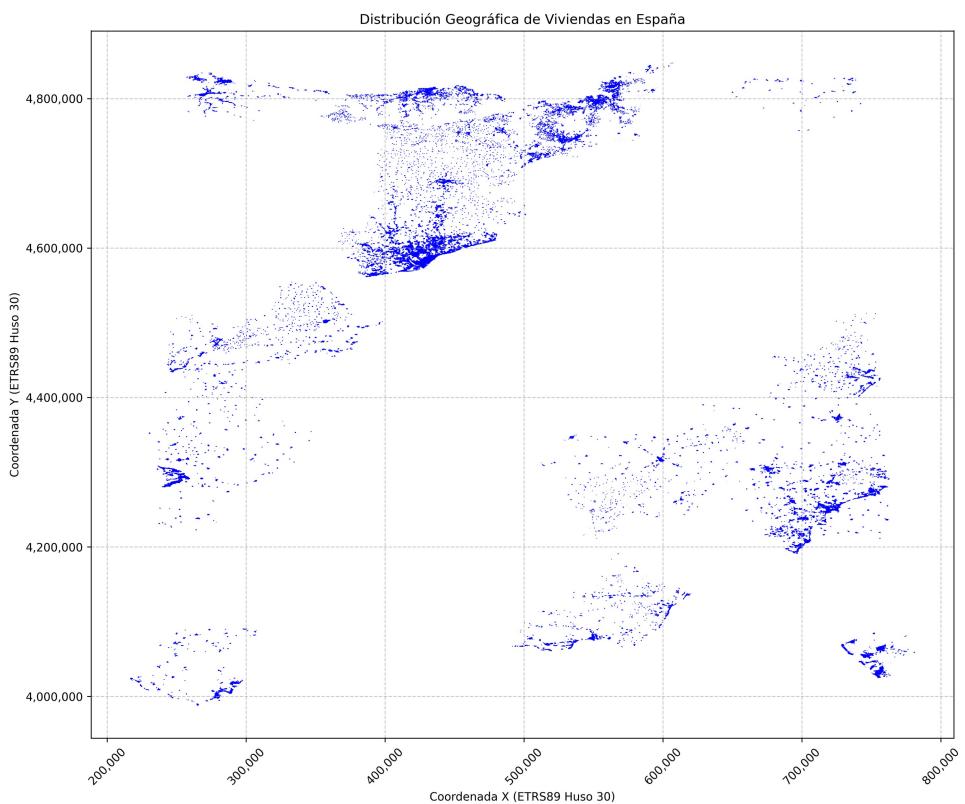


Figura 23: Distribución geográfica de los datos de viviendas catastrales en España utilizados para el Escenario 1 (Fuente: Dirección General del Catastro, procesado por el proyecto). Elaboración del autor.

Cuando los datos catastrales detallados están disponibles y habilitados para una provincia, el modelo utiliza el archivo `viviendas_2014_2020.csv` (producto del subproyecto CATASTRO). Este archivo contiene información como la referencia catastral, coordenadas del centroide del edificio, superficie y año de alta. El método `_create_houses_real` en `world/generator.py` se encarga de procesar estos datos. La asignación de estas viviendas a un municipio específico dentro del modelo se realiza mediante operaciones espaciales estándar, como la de punto en polígono, asegurando que cada vivienda catastral se asocie correctamente a su municipio correspondiente. Este enfoque proporciona la máxima fidelidad en cuanto a la ubicación y características básicas de las viviendas nuevas incorporadas al modelo durante el período cubierto por los datos.

Escenario 2: Generación Sintética de Viviendas. En ausencia de datos catastrales utilizables para una provincia, o si así se configura, el modelo activa el método `_create_houses_synthetic`. Este proceso recurre a la geometría de los polígonos municipales para generar y ubicar viviendas de forma sintética pero espacialmente distribuida:

- **Carga de Polígonos Municipales:** Inicialmente, el sistema carga las geometrías de los límites territoriales de los municipios. En el código, la función `prepare_shapes(sim.geo)` carga estas geometrías, que luego se utilizan para crear un GeoDataFrame donde cada municipio tiene asociado su polígono real.
- **Triangulación del Polígono Municipal:** Para distribuir las viviendas sintéticas de manera realista dentro de los confines de un municipio, en lugar de un simple punto aleatorio en un bounding box, el modelo primero realiza una triangulación del polígono municipal. Utilizando la librería Shapely (específicamente, su función `triangulate`), tanto la geometría del área urbana del municipio como la geometría completa del municipio se descomponen en una serie de triángulos no solapantes. Las Figuras 24 y 25 muestran ejemplos visuales de este proceso de triangulación aplicado a polígonos municipales.

- **Generación de Ubicaciones Aleatorias Ponderadas:** Una vez triangulado el polígono, se calcula el área de cada triángulo. Para generar una ubicación sintética, primero se selecciona un triángulo de forma probabilística, ponderando la elección por el área de cada triángulo (los triángulos más grandes tienen más probabilidad de ser elegidos). Luego, se genera un punto aleatorio dentro del triángulo seleccionado utilizando coordenadas baricéntricas. Este método asegura una distribución de viviendas más homogénea y representativa de la extensión territorial del municipio. La Figura 26 ilustra la distribución de puntos (viviendas) generados mediante este método en dos municipios.
- **Fallback al Centroide:** En casos donde la triangulación pueda fallar (por ejemplo, debido a geometrías inválidas o áreas nulas), el sistema recurre a utilizar el centroide del polígono municipal como ubicación para las viviendas sintéticas.

Este enfoque dual permite al modelo PolicySpace2 adaptado a España operar con el mayor nivel de detalle espacial posible, utilizando datos catastrales reales siempre que sea factible, y recurriendo a un método de generación sintética espacialmente informado cuando es necesario.

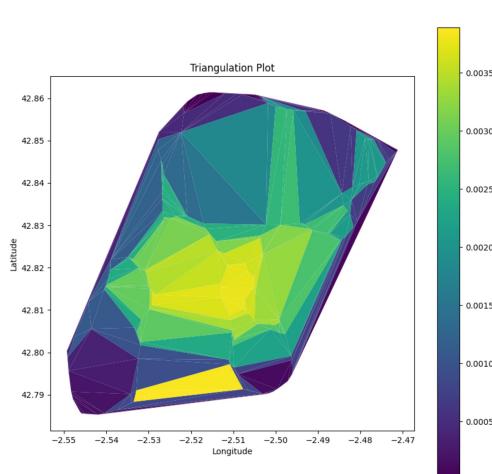


Figura 24: Ejemplo de triangulación de un polígono municipal (01001). Elaboración del autor.

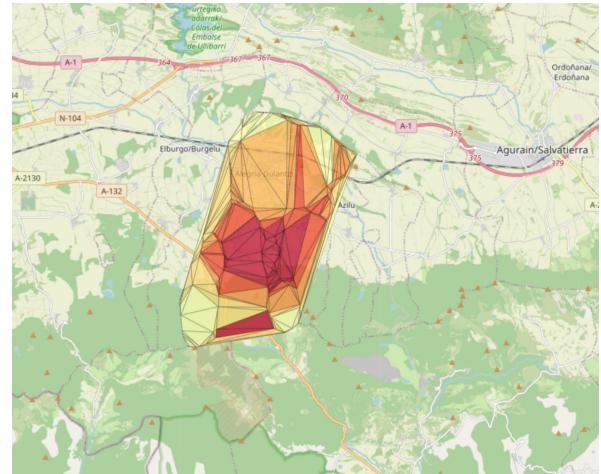


Figura 25: Otro ejemplo de triangulación de un polígono municipal superpuesto a un mapa base. Elaboración del autor.

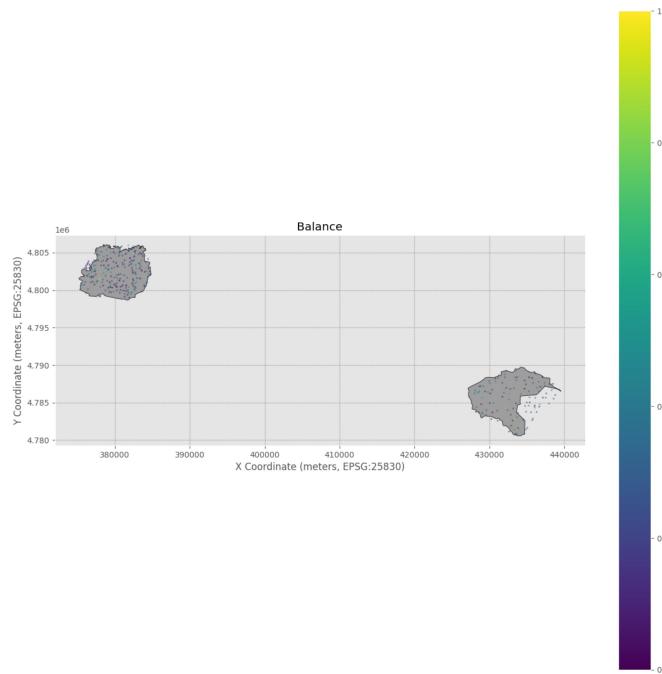


Figura 26: Visualización de puntos (viviendas sintéticas) generados dentro de los polígonos de dos municipios tras el proceso de triangulación. Elaboración del autor.

La implementación detallada de estos mecanismos se encuentra en el script `world/generator.py` dentro del repositorio del proyecto `PolicySpace2_Spain_new_ETL`.

5.11.3. Modelo de Precios de Vivienda (proyecto_modelo_1)

El proyecto `proyecto_modelo_1` tuvo como objetivo analizar y predecir los precios del mercado inmobiliario español y extraer las principales variables de las que disponemos. Se utilizaron datos de Kaggle (Spanish Housing Dataset) y se exploró su integración con el `dataset_municipio_cnae_anual_2014_2020.csv`. El notebook de referencia empleado fué el propuesto por el usuario Martin Lopez15, `spanish-housing.ipynb` detalla el preprocesamiento de los datos de Kaggle y la implementación de un modelo de Red Neuronal con TensorFlow/Keras para predecir el precio por metro cuadrado. El script `notebooks/housing_model_integrated.py` muestra cómo se integraron los datos de vivienda de Kaggle con el dataset municipal socioeconómico. Posteriormente, el script `notebooks/modelo_predictivo_corregido.py` implementó y se compara varios modelos de machine learning (Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost) utilizando el dataset integrado, realizando análisis de importancia de variables y validación de resultados.

El código fuente de este subproyecto, junto con los notebooks y scripts detallados, está disponible en el repositorio de GitHub: https://github.com/agmalaga2020/proyecto_modelo_1. Además, se ha desarrollado una demostración interactiva de los resultados y análisis de este modelo predictivo, accesible en: https://agmalaga2020.github.io/proyecto_modelo_1/web/.

5.11.4. Resultados de Simulación y Documentación Adicional

Los resultados de las simulaciones del modelo PolicySpace2 adaptado se presentan en una infografía HTML ([PolicySpace2_Spain_new_ETL/post_analysis/.../infografia_comparativa_politicas.html](#)). Esta infografía muestra tendencias generales agregadas (commuting, Gini, QLI, alquiler, desempleo), análisis a nivel municipal para un conjunto de municipios seleccionados, comparativas puntuales y mapas coropléticos.

6. Planificación y calendario

Para la gestión del proyecto, se utilizará la plataforma Monday, la cual permite organizar y monitorizar eficientemente el progreso de las tareas. Nos enfocaremos en dos vistas principales:

- **Vista de tabla** : Esta vista permite organizar las tareas, asignar responsables, establecer prioridades y definir dependencias. Se utiliza como base para controlar el estado de cada tarea y realizar ajustes según las necesidades del proyecto.

Tarea	Responsable	Estado	Fecha	Prioridad	Notas	Archivos	Cronograma	Última actualizac...	Depende de
Investigación inicial del sector inmobiliario	AG	En curso	15 mar.	Alta	Recopilación de datos sobre tendencias del mercado		15 - 20 mar.	Hace 13 m...	
Identificación de fuentes de datos	AG	No iniciado	21 mar.	Alta	Fuentes oficiales y privadas		21 - 25 mar.	Hace 9 m...	Investigación inicial ...
Creación del pipeline ETL	AG	No iniciado	26 mar.	Alta	Extracción, Transformación y Carga de datos		26 - 31 mar.	Hace 9 m...	Identificación de fue...
Configuración de infraestructura tecnológica	AG	No iniciado	1 abr.	Alta	Configuración de AWS y herramientas necesarias		1 - 5 abr.	Hace 6 m...	Creación del pipeline...
Desarrollo del modelo ABM	AG	No iniciado	6 abr.	Alta	Creación de simulaciones básicas en agentes		6 - 20 abr.	Hace 8 m...	Configuración de inf...
Creación del repositorio en GitHub	AG	No iniciado	21 abr.	Media	Organización y documentación del código del proy...		21 - 23 abr.	Hace 6 m...	Configuración de inf...
Implementación de modelos predictivos	AG	No iniciado	24 abr.	Media	Entrenamiento de modelos de ML para predicciones		24 - 30 abr.	Hace 8 m...	Desarrollo del modell...
Evaluación de políticas públicas	AG	No iniciado	1 may.	Alta	Comparativa de impacto social y económico de polí...		1 - 10 may.	Hace 6 m...	Implementación de ...
Diseño de visualizaciones	AG	No iniciado	11 may.	Media	Creación de gráficos, mapas y dashboards		11 - 20 may.	Hace 6 m...	Evaluación de polític...
Validación de resultados	AG	No iniciado	21 may.	Alta	Revisión de precisión, impacto y coherencia		21 - 31 may.	Hace 6 m...	Diseño de visualizaci...
Redacción del informe final	AG	No iniciado	1 jun.	Alta	Elaboración y revisión de las conclusiones		1 - 10 jun.	Hace 5 m...	Validación de resulta...
Presentación del proyecto	AG	No iniciado	11 jun.	Alta	Preparación de diapositivas y entrega final		11 - 15 jun.	Hace 5 m...	Redacción del informe...
Creación del esquema relacional de la BD	AG	No iniciado	15 abr.	Alta	Diseño del modelo lógico y físico		15 - 20 abr.	Hace 5 m...	Creación del pipeline...
Optimización de scraper para fuentes inmobiliarias	AG	No iniciado	5 may.	Media	Ajustes para extracción eficiente de datos		5 - 8 may.	Hace 5 m...	Identificación de fue...
Análisis geoespacial de datos	AG	No iniciado	9 may.	Media	Uso de geopandas para analizar dinámicas espaciales		9 - 12 may.	Hace 5 m...	Configuración de inf...
Creación del pipeline ETL	AG	No iniciado						Hace 6 m...	

Figura 27: Vista de tabla. Elaboración del autor.

- **Vista de cronograma Gantt** : Esta vista ofrece una representación temporal del proyecto, facilitando la visualización de tareas en una línea de tiempo. Es especialmente útil para gestionar dependencias entre actividades y ajustar plazos ante posibles retrasos o cambios en el alcance.

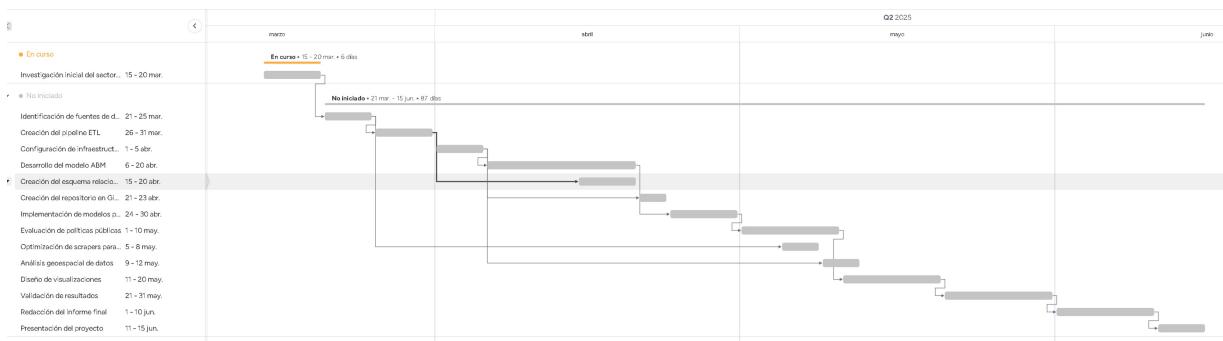


Figura 28: Vista Gantt. Elaboración del autor.

El proyecto tiene una duración aproximada de tres meses, desde el **15 de marzo** hasta el **15 de junio**. Las tareas clave incluyen:

- Investigación inicial y recopilación de datos.
- Desarrollo del proceso ETL y su implementación.
- Configuración del repositorio en GitHub para la colaboración del equipo.
- Implementación de modelos basados en machine learning.
- Diseño y creación de visualizaciones para la presentación de resultados.
- Redacción y entrega del informe final.

Con estas herramientas y vistas, se garantiza una planificación clara, el seguimiento del progreso en tiempo real y la capacidad de adaptarse a imprevistos durante el desarrollo del proyecto.

7. Análisis y Gestión de Riesgos

La implementación de este proyecto presenta diversos riesgos que deben identificarse, evaluarse y gestionarse para garantizar el éxito del trabajo. A continuación, se detallan los principales riesgos, su probabilidad de ocurrencia, impacto potencial y las estrategias para mitigarlos.

7.1. Identificación de Riesgos

■ Riesgos relacionados con los datos:

- Acceso limitado a datos inmobiliarios reales debido a restricciones legales o costos elevados.
- Calidad de los datos obtenidos mediante *web scraping*, que puede incluir información incompleta o inexacta.
- Incompatibilidad entre diferentes fuentes de datos que dificulten su integración.

■ Riesgos técnicos:

- Limitaciones en la infraestructura tecnológica para manejar grandes volúmenes de datos.
- Complejidad en la implementación de simulaciones basadas en agentes ([ABM](#)).
- Sobreajuste o falta de precisión en los modelos de [Machine Learning \(ML\)](#).

■ Riesgos legales y éticos:

- Incumplimiento del Reglamento General de Protección de Datos (RGPD) al manejar datos sensibles.
- Posibilidad de sesgos algorítmicos que afecten negativamente la equidad de las soluciones propuestas.

■ Riesgos organizativos:

- Falta de coordinación en la integración de múltiples herramientas y plataformas.
- Retrasos en la ejecución debido a dificultades técnicas o falta de recursos.

7.2. Evaluación de Probabilidad e Impacto

Riesgo	Probabilidad	Impacto	Estrategia de Mitigación
Acceso limitado a datos inmobiliarios	Alta	Alto	Buscar proxies y fuentes alternativas como TINSA o datos abiertos oficiales.
Calidad de datos obtenidos mediante <i>web scraping</i>	Media	Alto	Implementar procesos de validación y limpieza de datos robustos.
Limitaciones tecnológicas	Media	Alto	Configurar infraestructura en la nube (e.g., AWS) y optimizar código para manejo de datos masivos.
Sesgos algorítmicos	Baja	Alto	Monitorear modelos para identificar y corregir posibles sesgos.
Incumplimiento del RGPD	Baja	Muy Alto	Implementar anonimización de datos y consultar con expertos legales.
Falta de coordinación entre herramientas	Media	Medio	Utilizar herramientas de gestión de proyectos y documentar procesos claramente.

Cuadro 2: Evaluación de riesgos con estrategias de mitigación.

7.3. Plan de Gestión de Riesgos

Para minimizar los riesgos identificados, se adoptarán las siguientes acciones:

- **Monitoreo continuo:** Establecer un plan de seguimiento regular para evaluar la calidad de los datos, el desempeño de los modelos y el cumplimiento normativo.
- **Capacitación técnica:** Invertir tiempo en la capacitación del equipo en herramientas específicas como simulaciones basadas en agentes y [ETL](#).
- **Consultoría legal:** Consultar con expertos en normativa de protección de datos y mercado inmobiliario para asegurar el cumplimiento legal y ético.
- **Pruebas preliminares:** Realizar pruebas iniciales con datos de menor escala para identificar posibles problemas técnicos o algorítmicos antes de pasar a un entorno más amplio.

Este análisis y gestión de riesgos aseguran que el proyecto se desarrolle de manera eficiente, minimizando impactos negativos y maximizando el logro de los objetivos establecidos.

8. Herramientas y Fuentes para la Construcción de la Base del Proyecto

Para construir una base sólida en el proyecto, se emplean y emplearán fuentes académicas y profesionales relevantes. A continuación, se describen las principales herramientas y plataformas utilizadas para su búsqueda y organización:

8.1. Connected Papers

Esta plataforma permite acceder a documentos científicos conectados por rama de investigación, facilitando la profundización en apartados específicos. Es una herramienta clave para explorar y comprender los vínculos entre investigaciones relacionadas.

8.2. Mendeley

Se utiliza como pilar principal para la organización de documentos académicos. Ofrece una forma sencilla de gestionar estudios y referencias, permitiendo clasificarlos en carpetas, lo que resulta esencial para mantener el orden en las bases teóricas del proyecto.

8.3. Google Scholar

Proporciona un acceso amplio a documentos académicos, con una ventaja destacable en la facilidad para citar referencias. Sin embargo, tiene limitaciones en los filtros de relevancia y citas. Para superar esto, se está utilizando SerpApi, que mediante un pequeño código en [script](#) nos facilita la búsqueda de documentos relevantes a partir de palabras clave específicas.

8.4. Kaggle

Esta plataforma aporta soluciones técnicas prácticas y comparativas a problemas específicos. Permite analizar diferentes enfoques y explorar nuevas soluciones aplicadas a un mismo objetivo. Además, es una fuente rica en datos del sector inmobiliario, lo que resulta de gran valor para este proyecto.

8.5. VLex

Es una fuente estructurada en el ámbito legal, que proporciona acceso a legislación vigente y derogada, doctrinas, jurisprudencia, boletines oficiales, y otros recursos útiles para construir un marco legal y ético relacionado con la inteligencia artificial.

Aunque no se incluirá esta misma dentro de los procesos [ETL](#) debido a su coste, se está investigando con datos previamente filtrados en grandes volúmenes para realizar estudios preliminares de etiquetado. Estos datos se están organizando en una base de datos con variables categóricas codificadas generadas por un [LLM](#), como son:

- **Subvención:** 0 = No, 1 = Sí.
- **Regulación activa:** 0 = No, 1 = Sí.
- **Inversión pública:** 0 = No, 1 = Sí.
- **Impacto ambiental evaluado:** 0 = No, 1 = Sí.
- **Participación ciudadana:** 0 = No, 1 = Sí.

El modelo, está generando también etiquetas específicas, tales como:

- Cédula de Habitabilidad.
- Prohibición de Uso Turístico en Suelo Rústico.
- Medidas de Eficiencia Energética.
- Fomento del Alquiler de Viviendas Vacías.
- Moratoria de Plazas Turísticas.
- Gestión de Suelo Municipal.
- Rehabilitación de Edificios.
- Impacto de Precios de Alquiler.
- Cambio de Uso de Local a Vivienda.

8.6. GitHub

Complementa el trabajo técnico al proporcionar soluciones prácticas, como el desarrollo de scrapers para fuentes inmobiliarias y el acceso a bases de datos adicionales no disponibles en otras plataformas. Es una herramienta clave para compartir y colaborar en soluciones de código abierto aplicadas al proyecto.

Estas herramientas forman un conjunto robusto que permite abordar el proyecto con una base teórica sólida, acceso a datos relevantes y soluciones técnicas innovadoras, asegurando la calidad y profundidad del trabajo.

8.7. Fuentes de Datos Primarias y Organismos Oficiales

Una parte fundamental del proyecto se basa en la recopilación y procesamiento de datos provenientes de organismos oficiales, que proporcionan la información base para la calibración y validación del modelo:

- **Instituto Nacional de Estadística (INE):** Principal proveedor de datos demográficos (población, mortalidad, fecundidad, tamaño de hogares), económicos (empresas, nivel educativo, PIE) y territoriales para España.
- **Dirección General del Catastro:** Fuente de los datos catastrales de edificaciones, crucial para la representación detallada de las viviendas en el modelo.
- **Banco de España y Banco Central Europeo (BCE):** Proveedores de datos financieros, como las tasas de interés hipotecarias y oficiales.
- **Eurostat:** Fuente de indicadores económicos a nivel europeo, como el Índice de Precios de Consumo Armonizado (IPCA) utilizado para calcular tasas de interés reales.
- **Centro Nacional de Información Geográfica (CNIG):** Proporciona datos geoespaciales, como la superficie municipal, utilizados en diversos análisis.
- **Ministerio de Hacienda y otros ministerios:** Fuentes de datos fiscales y de políticas públicas específicas.

8.8. Entorno de Desarrollo y Lenguajes de Programación

El desarrollo técnico del proyecto se apoya en un conjunto de tecnologías de programación y análisis de datos:

- **Python:** Lenguaje principal para el desarrollo de los scripts ETL, los modelos de simulación (PolicySpace2), los modelos de machine learning y los dashboards.
- **Librerías de Python:**
 - *Pandas y Polars:* Para la manipulación y análisis eficiente de grandes volúmenes de datos tabulares.
 - *NumPy:* Para operaciones numéricas.
 - *Scikit-learn:* Para la implementación de modelos de machine learning (regresión, clasificación, clustering).
 - *TensorFlow y Keras:* Para el desarrollo de modelos de redes neuronales profundas, específicamente en la predicción de precios de vivienda.
 - *GeoPandas y Shapely:* Para el tratamiento de datos geoespaciales, manipulación de polígonos y cálculos espaciales.
 - *Matplotlib y Seaborn:* Para la generación de gráficos y visualizaciones estáticas en los análisis exploratorios y de resultados.
 - *Requests, BeautifulSoup (o similares):* Potencialmente utilizadas para tareas de web scraping en la recopilación inicial de ciertos datos o referencias.
- **Jupyter Notebooks/JupyterLab:** Entorno interactivo para el desarrollo y documentación de los procesos ETL y análisis exploratorios.

8.9. Sistemas de Gestión de Bases de Datos

Para el almacenamiento y gestión de los datos procesados, se han empleado:

- **SQLite:** Utilizada para la base de datos inicial del proyecto (`datawarehouse.db`) y la base de datos mejorada (`base_datos_mejorada.db`), facilitando la portabilidad y el acceso local.
- **PostgreSQL con PostGIS (mencionado como alternativa):** Considerado para una gestión más robusta de datos geoespaciales a gran escala, aunque la implementación actual se basa en SQLite.

8.10. Herramientas de Visualización, BI y Documentación

- **Streamlit:** Para la creación de dashboards interactivos que permiten la exploración de los datos procesados y los resultados de las simulaciones.
- **Graphviz y Eralchemy:** Utilizadas para generar esquemas visuales de las bases de datos del proyecto.
- **Power BI:** Mencionado como herramienta utilizada para la creación de visualizaciones de ejemplo (ej. ‘datos ejemplo power bi.png’).
- **LaTeX (Overleaf):** Para la redacción colaborativa y profesional de la documentación del proyecto, incluyendo este mismo documento.

8.11. Herramientas de Gestión de Proyectos y Colaboración

- **Monday.com:** Plataforma utilizada para la planificación, seguimiento de tareas y gestión general del proyecto.
- **Git:** Sistema de control de versiones utilizado implícitamente a través de GitHub para la gestión del código fuente y la colaboración.
- **Notion:** Herramienta utilizada para la organización de ideas, documentación y seguimiento colaborativo de aspectos del proyecto.

9. Implicaciones para el Diseño del Proyecto

9.1. Ética y Legalidad

El diseño del proyecto debe garantizar el cumplimiento de las regulaciones legales, como el [RGPD](#), y evitar posibles violaciones a la privacidad de los datos. Es fundamental limitar el [scraping](#) y el análisis a información no sensible, implementando medidas para anonimizar y proteger la información procesada.

9.2. Limitaciones Técnicas

- **Calidad de datos:** Los datos extraídos de fuentes como Idealista pueden requerir procesos robustos de limpieza y normalización para corregir errores y asegurar su representatividad.
- **Procesamiento masivo:** Será necesario configurar una infraestructura tecnológica adecuada, como AWS, para procesar y analizar los datos de manera eficiente y escalable.
- **Anonimización:** Implementar estrategias para garantizar que ningún dato recopilado permita identificar a personas o entidades, minimizando riesgos legales y éticos.

9.3. Líneas de Acción

- **Transparencia:** Documentar los procesos de recopilación y análisis de datos para demostrar conformidad con normativas éticas y legales.
- **Tecnología:** Diseñar flujos de trabajo que incluyan herramientas como ETL y AWS para un procesamiento eficiente de los datos.

10. Resultados del Proyecto y Discusión

10.1. Contexto Analítico y Hallazgos de la Literatura Previa

10.1.1. Riesgos Éticos/Legales frente a Beneficios Sociales de la IA en Vivienda

Aspecto	Riesgos Éticos/Legales	Beneficios Sociales
Acceso a la Vivienda	Exclusión social por sesgos en modelos predictivos y precios	Mejora en la accesibilidad al identificar oportunidades de vivienda asequible
Privacidad de Datos	Riesgo de exposición de datos sensibles de propietarios e inquilinos	Optimización en la planificación urbana y distribución justa del suelo
Gentrificación	Potenciación de la gentrificación por plataformas de alquiler como Airbnb	Regulación más eficiente basada en análisis predictivo para evitar desplazamientos

Cuadro 3: Comparativa de Riesgos Éticos/Legales frente a Beneficios Sociales de la IA en Vivienda.

10.1.2. Comparativa de Técnicas de ML Aplicadas al Sector Vivienda

Técnica	Precisión	Interpretabilidad	Escalabilidad
Redes Neuronales	Alta para predicciones de precios complejos	Baja, limitada para reguladores urbanos	Alta en grandes conjuntos de datos, como portales inmobiliarios
Árboles de Decisión	Buena para decisiones regulatorias	Alta, útil para analizar dinámicas locales	Limitada a nivel nacional sin ajustes
Regresiones Lineales	Moderada para tendencias macroeconómicas	Alta, más sencilla para planes de políticas	Escalabilidad adecuada para datasets medianos

Cuadro 4: Comparativa de Técnicas de Machine Learning Aplicadas al Sector Vivienda.

10.1.3. Hallazgos Clave de la Literatura

- **Impactos Técnico-Sociales:** Las plataformas digitales como Airbnb modifican las dinámicas del mercado, exacerbando la exclusión. Sin embargo, la IA puede ayudar a mitigar estos efectos mediante herramientas predictivas y explicativas.
- **Regulación Basada en Datos:** El uso de técnicas de Machine Learning (ML) permite planificar políticas que reduzcan desigualdades y mejoren el acceso a la vivienda.
- **Convergencias:** La combinación de modelos explicativos, como los árboles de decisión, y modelos predictivos, como redes neuronales, optimiza tanto la planificación como la ejecución de políticas públicas.

10.2. Resultados del Modelo Predictivo de Precios de Vivienda

El desarrollo de un modelo predictivo de precios de vivienda fue un componente clave del proyecto, utilizando tanto datos del portal Kaggle (Spanish Housing Dataset) como datos socioeconómicos municipales integrados. Se exploraron diversos enfoques, desde redes neuronales hasta modelos clásicos de machine learning, con el objetivo de comprender los factores determinantes del precio y generar predicciones robustas.

10.2.1. Rendimiento y Comparación de Modelos

En una primera aproximación, documentada en el análisis del notebook `spanish-housing.ipynb` (referenciado como `spanish-housing-analysis.md`), propiedad del autor Martin Lopez15, se empleó un modelo de Red Neuronal con TensorFlow/Keras sobre el dataset de Kaggle. Este dataset, centrado en características de las viviendas, fue sometido a un preprocesamiento que incluyó el tratamiento de outliers, la codificación de variables categóricas y el escalado de características. El modelo de red neuronal, al predecir el logaritmo del precio por metro cuadrado, alcanzó un Error Cuadrático Medio (MSE) de aproximadamente 0.0088 y un Error Absoluto Medio (MAE) de 0.067 en el conjunto de prueba. Estas métricas iniciales indicaron una capacidad predictiva prometedora.

Posteriormente, el script `modelo_predictivo_corregido.py` (ubicado en `fases del proyecto`) propuso una ampliación del análisis evaluando un conjunto de modelos de regresión más tradicionales sobre un dataset Enriquecido. Este dataset integraba las características de las viviendas del dataset de Kaggle con variables socioeconómicas a nivel municipal, provenientes del archivo `housing_municipal_integrated.fixed.csv`. Los modelos probados incluyeron Regresión Ridge, Regresión Lasso, Random Forest, Gradient Boosting y XGBoost.

Las comparaciones de rendimiento se realizaron utilizando dos configuraciones de características:

1. **Solo variables de vivienda:** Utilizando únicamente las características intrínsecas de las propiedades y su ubicación (provincia, tipo de vivienda).
2. **Variables de vivienda + socioeconómicas:** Combinando el conjunto anterior con indicadores municipales como población, Índice de Desarrollo Humano (IDH) y número de empresas.

La Tabla 5 resume las métricas de rendimiento (RMSE, MAE y R^2) para los modelos evaluados en ambos conjuntos de características. Consistentemente, los modelos basados en ensambles de árboles (Random Forest, Gradient Boosting, XGBoost) mostraron un rendimiento superior a los modelos lineales (Ridge, Lasso).

Cuadro 5: Comparación del rendimiento de modelos predictivos de precios de vivienda.

Modelo	Dataset	RMSE	MAE	R^2
Ridge	Solo Vivienda	277234.72	162778.50	0.5360
Lasso	Solo Vivienda	277234.49	162778.01	0.5360
RandomForest	Solo Vivienda	100307.98	58894.69	0.8670
GradientBoosting	Solo Vivienda	108030.09	66080.93	0.8459
XGBoost	Solo Vivienda	99580.06	59678.77	0.8693
Ridge	Con Socioeconómicas	240490.83	144484.48	0.6632
Lasso	Con Socioeconómicas	240487.33	144480.89	0.6632
RandomForest	Con Socioeconómicas	80004.10	48903.59	0.9157
GradientBoosting	Con Socioeconómicas	84732.11	52639.97	0.9054
XGBoost	Con Socioeconómicas	78306.23	48619.96	0.9235

Un hallazgo crucial fue el impacto positivo de la inclusión de variables socioeconómicas. Para todos los modelos, añadir estos predictores contextuales resultó en una mejora significativa de las métricas de evaluación. Como se observa en la Figura 29 y la Figura 30, el modelo XGBoost con variables socioeconómicas alcanzó el mejor desempeño general (R^2 de 0.9235 y RMSE de 78306.23), lo que subraya la importancia de los factores contextuales municipales en la predicción de los precios de la vivienda.

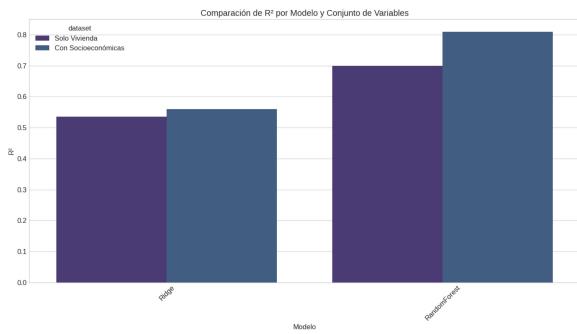


Figura 29: Comparación del R^2 entre modelos. Elaboración del autor.

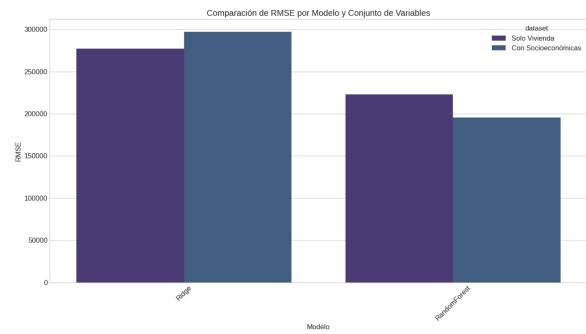


Figura 30: Comparación del RMSE entre modelos. Elaboración del autor.

10.2.2. Análisis de Variables Más Importantes

El análisis de la importancia de las variables, realizado sobre el modelo XGBoost con el conjunto completo de características (mostrado en `modelo_predictivo_corregido.py`), reveló los factores más determinantes del precio de la vivienda. La Figura 31 muestra las 20 variables más influyentes.

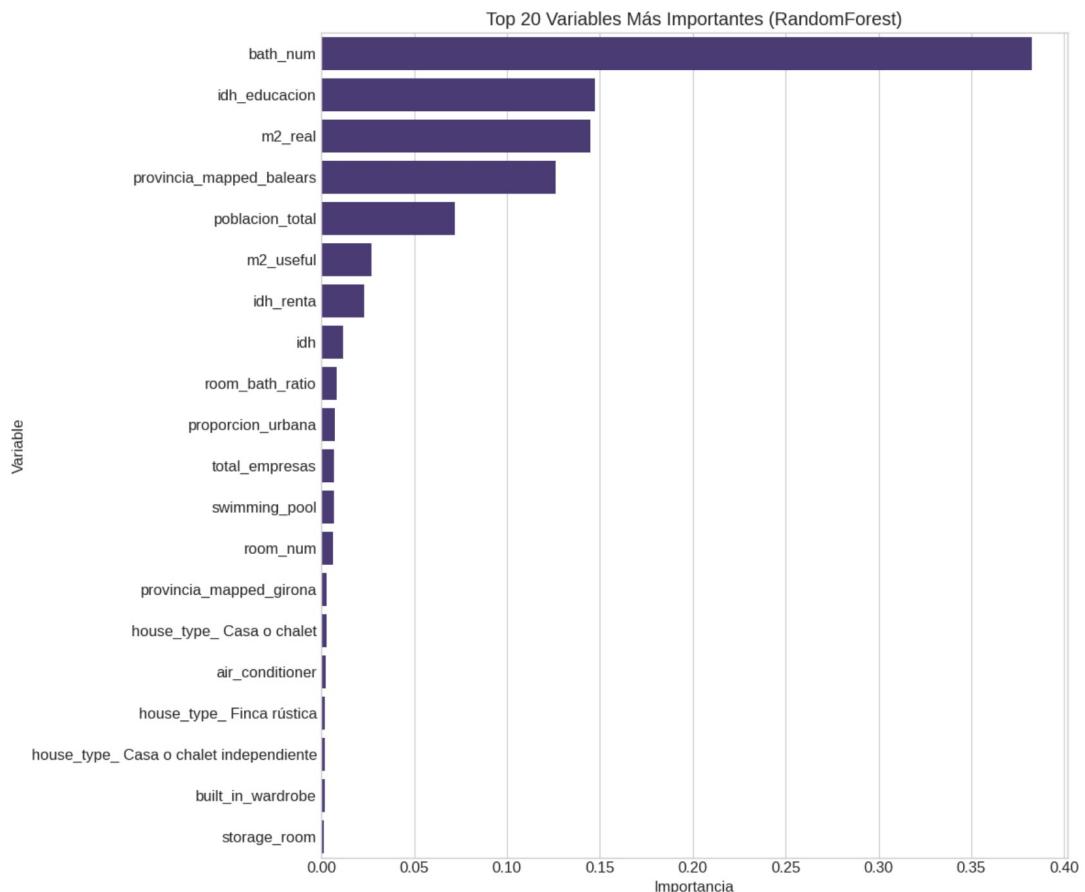


Figura 31: Las 20 variables más importantes según el modelo XGBoost. Elaboración del autor.

Las variables que consistentemente mostraron mayor poder predictivo fueron:

- **Características intrínsecas de la vivienda:** La superficie real (`m2_real`) y la superficie útil (`m2_useful`) emergieron como los predictores más potentes, seguidas por el número de baños (`bath_num`) y el número de habitaciones (`room_num`).
- **Comodidades y extras:** La presencia de ascensor (`lift`), armarios empotrados (`built_in_wardrobe`) y aire acondicionado (`air_conditioner`) también figuraron entre las variables relevantes.
- **Ubicación y tipo de propiedad:** La provincia (`provincia_mapped`, codificada) y el tipo de vivienda (`house_type`, codificada) demostraron ser significativas, reflejando las disparidades geográficas y las diferencias de valoración entre tipologías.
- **Variables socioeconómicas municipales:** El Índice de Desarrollo Humano (`idh`) del municipio y la población total (`poblacion_total`) fueron las variables socioeconómicas más destacadas, confirmando su influencia en los precios.

Al agrupar la importancia relativa de los diferentes tipos de variables (Figura 32), se constató que las características propias de la vivienda aportaban la mayor parte de la capacidad explicativa del modelo. No obstante, las variables categóricas (especialmente la ubicación) y las socioeconómicas, aunque con un peso individual menor, resultaron cruciales para alcanzar los niveles más altos de precisión predictiva.

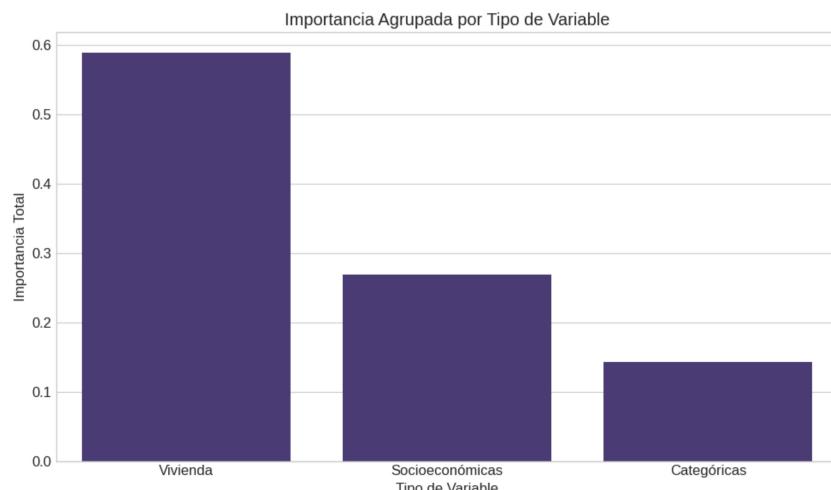


Figura 32: Importancia de las variables agrupadas por tipo. Elaboración del autor.

Las Figuras 33 y 34 ilustran la relación del precio con dos de las variables más importantes: la superficie y el IDH municipal, respectivamente, mostrando tendencias visuales que el modelo captura.

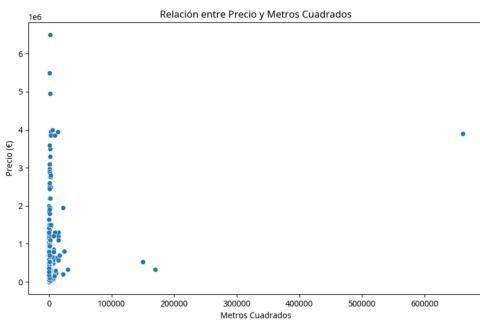


Figura 33: Relación entre precio y metros cuadrados.
Elaboración del autor.

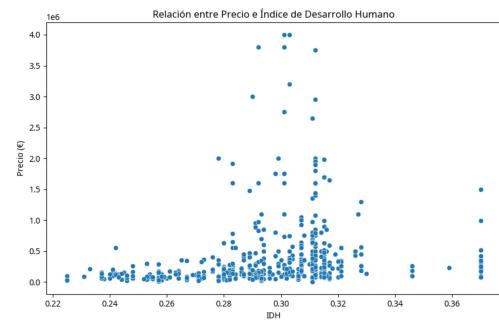


Figura 34: Relación entre precio e IDH municipal.
Elaboración del autor.

10.2.3. Validación del Modelo

La validación de los modelos predictivos se abordó mediante varias estrategias. En el script, se utilizó una división estándar de los datos en conjuntos de entrenamiento (80 %) y de prueba (20 %). La optimización de hiperparámetros para cada modelo se realizó mediante búsqueda en rejilla con validación cruzada de 5 folds (`GridSearchCV`), utilizando el error cuadrático medio negativo como métrica de puntuación. Para el mejor modelo general obtenido (XGBoost con todas las características), se llevó a cabo un análisis de residuos para evaluar su bondad de ajuste. La Figura 35 muestra la distribución de los residuos, que se aproxima a una normal, mientras que la Figura 36 grafica los residuos frente a los valores predichos, no mostrando patrones sistemáticos evidentes (como heterocedasticidad), lo que sugiere una correcta especificación del modelo.

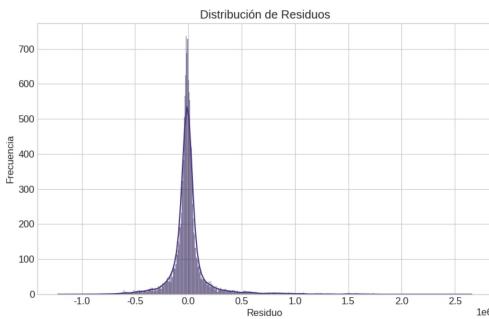


Figura 35: Distribución de los residuos del modelo XGBoost. Elaboración del autor.

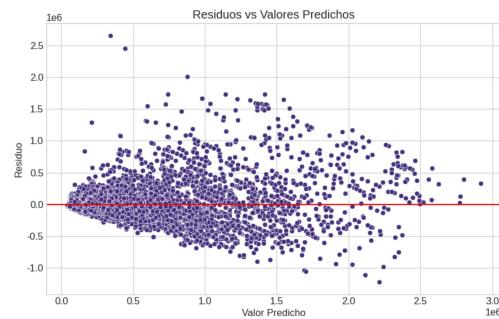


Figura 36: Gráfico de residuos vs. valores predichos.
Elaboración del autor.

Adicionalmente, el gráfico de dispersión de valores reales frente a valores predichos (Figura 37) mostró una fuerte concentración de puntos a lo largo de la diagonal de predicción perfecta, confirmando que las predicciones del modelo eran consistentes y precisas en relación con los precios observados, especialmente para el rango de precios más común, aunque con mayor dispersión en los valores más altos.

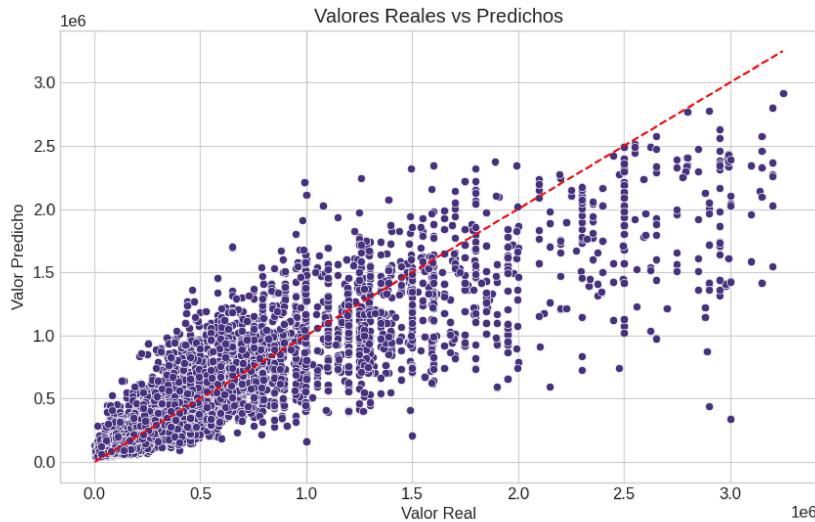


Figura 37: Comparación de valores reales frente a valores predichos por el modelo XGBoost. La línea discontinua roja representa la predicción perfecta. Elaboración del autor.

En el caso del modelo inicial de Red Neuronal, descrito en `spanish-housing-analysis.md`, la validación incluyó el seguimiento de las curvas de pérdida (MSE) durante las épocas de entrenamiento, observándose una convergencia adecuada tanto en los datos de entrenamiento como en los de validación, lo que sugiere que el modelo generalizó bien y no sufrió un sobreajuste excesivo. La comparación gráfica de valores predichos versus reales también mostró un ajuste razonable para este modelo.

10.3. Resultados de las Simulaciones con PolicySpace2 (Adaptación Española)

Las simulaciones realizadas con la adaptación española del modelo PolicySpace2, cuyos resultados se exploran en la infografía interactiva [infografia_comparativa_politicas.html](#) y en los gráficos individuales generados (disponibles en `_spain_2_POLICIES/plots_postanalysis/`), ofrecen una visión de las dinámicas socioeconómicas y del mercado inmobiliario bajo diferentes escenarios.

10.3.1. Escenario Base: Tendencias Agregadas y Dinámicas Municipales

Introducción al Escenario Base de Simulación

El análisis de los resultados de las simulaciones con el modelo PolicySpace2 adaptado al contexto español comienza con el estudio detallado del **escenario base**. Este escenario representa la evolución del sistema socioeconómico y del mercado inmobiliario sin la introducción de nuevas políticas públicas específicas más allá de las ya implícitas en la configuración inicial del modelo y los datos históricos utilizados para su calibración. La comprensión profunda de este escenario de referencia es fundamental por varias razones: primero, establece una línea de base contra la cual se podrán medir y atribuir los efectos de las intervenciones políticas que se analizarán posteriormente; segundo, permite observar las dinámicas “naturales” o intrínsecas del modelo, revelando tendencias, interacciones y posibles puntos de tensión que emergen de la propia estructura del sistema simulado y del comportamiento de los agentes.

En esta subsección, nos adentraremos en los múltiples resultados generados por el escenario base. Comenzaremos presentando las **tendencias agregadas** para el conjunto de los seis municipios de estudio simulados: **Valdemorales (10201)**, **Aín (12002)**, **Fisterra (15037)**, **Boal (33007)**, **Ramales de la Victoria (39057)** y **Voto (39102)**. Estas tendencias abarcan indicadores macroeconómicos y sociales clave, tales como la distancia media de commuting (desplazamiento diario al trabajo), el índice de Gini (como

medida de desigualdad de ingresos), el Índice de Calidad de Vida (QLI), la evolución de los precios del alquiler y la tasa de desempleo. El seguimiento de estos indicadores a lo largo del tiempo simulado nos proporciona una panorámica de la salud general del sistema para este conjunto de municipios, permitiendo identificar si el modelo tiende hacia equilibrios, ciclos, o si presenta crecimientos o decrecimientos sostenidos en variables críticas. Las Figuras 42 y 43 (presentadas más adelante) agrupan visualmente estas tendencias para facilitar su interpretación conjunta.

Posteriormente, el análisis descenderá al **nivel municipal**, centrándose en los seis municipios de estudio ya mencionados. Para estos casos específicos, se examinará la evolución de variables como el precio de venta de las viviendas (Figura 44), la distancia de commuting (Figura 45), el Producto Interior Bruto (PIB) per cápita (Figura 46), la tasa de desempleo (Figura 47), el índice de Gini a nivel regional (Figura 48) y la dinámica poblacional (Figura 49). Este desglose es crucial porque las tendencias agregadas a menudo ocultan una considerable heterogeneidad territorial. Las características particulares de cada municipio —su estructura económica, demografía, ubicación geográfica y dotación de servicios— pueden conducir a trayectorias muy dispares, incluso bajo las mismas condiciones macroeconómicas generales.

Además de las tendencias temporales, se explorarán **comparativas puntuales** en el último periodo de simulación para estos municipios de estudio. Estas instantáneas permitirán analizar indicadores como los permisos de construcción otorgados, el estado de la tesorería municipal, y las relaciones entre variables clave como la población y el PIB, o el QLI y el valor medio de la vivienda (Figuras 50, 51, y 52). Finalmente, se hará una breve mención a los **mapas coropléticos** (Figuras 53 y 54) que buscan ofrecer una perspectiva espacial de algunos de estos indicadores, aunque se reconocerán las limitaciones visuales presentes en la infografía original desde la que se derivan.

El conjunto de estos análisis del escenario base no solo valida la coherencia interna del modelo adaptado, sino que también sienta las bases para una evaluación rigurosa del impacto potencial de diversas políticas públicas, tema que se abordará en las subsecciones subsiguientes. Comprender cómo evoluciona el sistema “sin intervención”, es el primer paso para diseñar intervenciones efectivas y anticipar sus consecuencias.

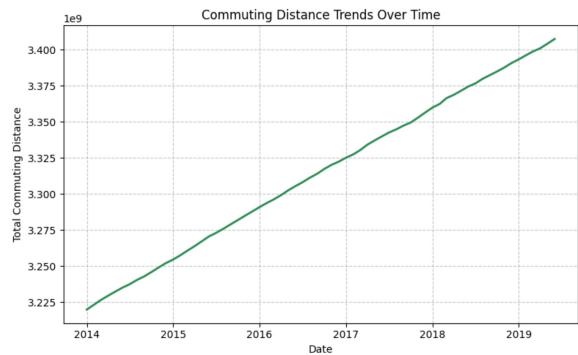


Figura 38: Tendencia de la distancia de commuting agregada. Elaboración del autor.



Figura 39: Tendencia del Índice de Gini agregado. Elaboración del autor.

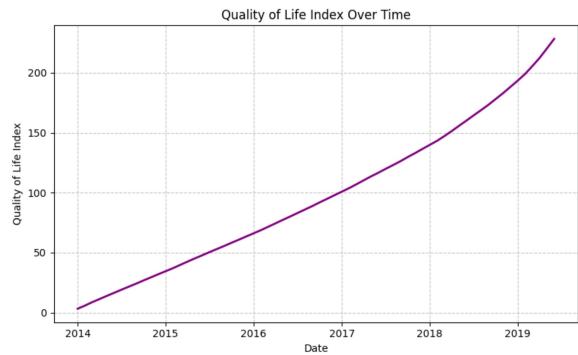


Figura 40: Tendencia del Índice de Calidad de Vida (QLI) agregado. Elaboración del autor.

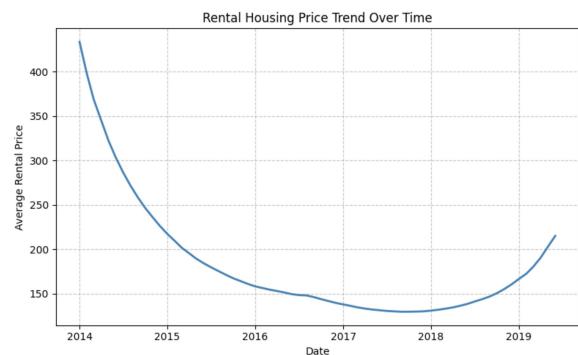


Figura 41: Tendencia del precio del alquiler agregado. Elaboración del autor.

Figura 42: Tendencias agregadas de indicadores socioeconómicos y de mercado en el escenario base (primeras cuatro métricas). Elaboración del autor.

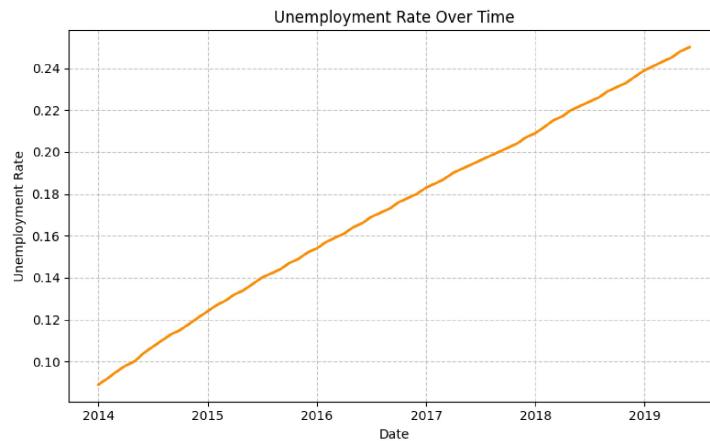


Figura 43: Tendencia de la tasa de desempleo agregada. Elaboración del autor.

A nivel municipal, para los seis municipios de estudio, se analizaron variables como el precio de venta de las viviendas, como vemos aún tenemos algún bug o problema de entrada de datos en el caso de Fiserra, color verde en (Figura 44), la distancia de commuting (Figura 45), el PIB per cápita (Figura 46), la tasa de desempleo (Figura 47), el índice de Gini regional (Figura 48) y la evolución de la población (Figura 49). Estas dinámicas variaron considerablemente, reflejando las particularidades de cada municipio.

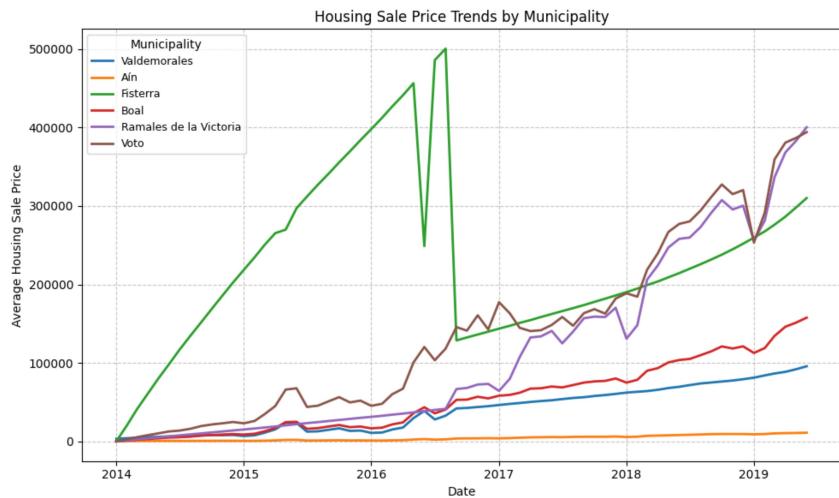


Figura 44: Evolución del precio de venta en municipios seleccionados. Elaboración del autor.

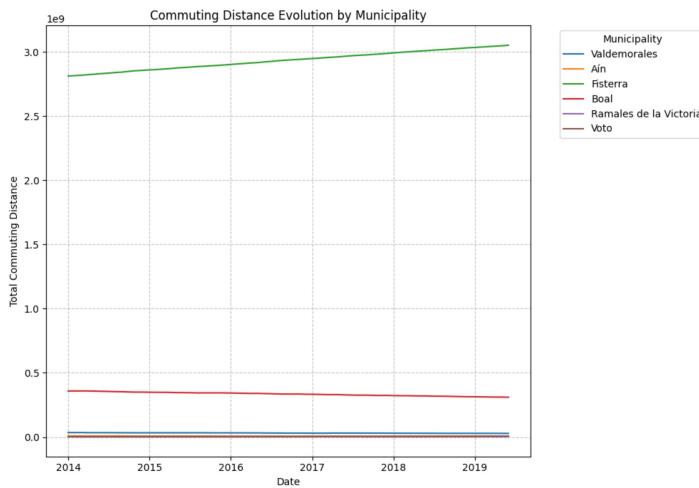


Figura 45: Evolución de la distancia de commuting en municipios seleccionados. Elaboración del autor.

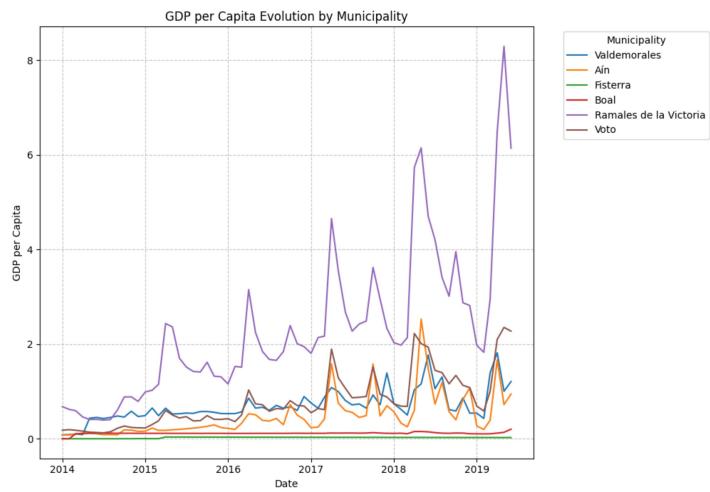


Figura 46: Evolución del PIB per cápita en municipios seleccionados. Elaboración del autor.

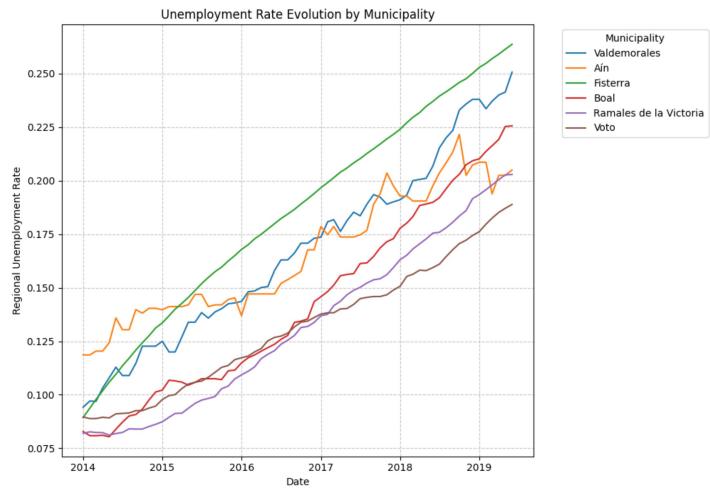


Figura 47: Evolución de la tasa de desempleo en municipios seleccionados. Elaboración del autor.

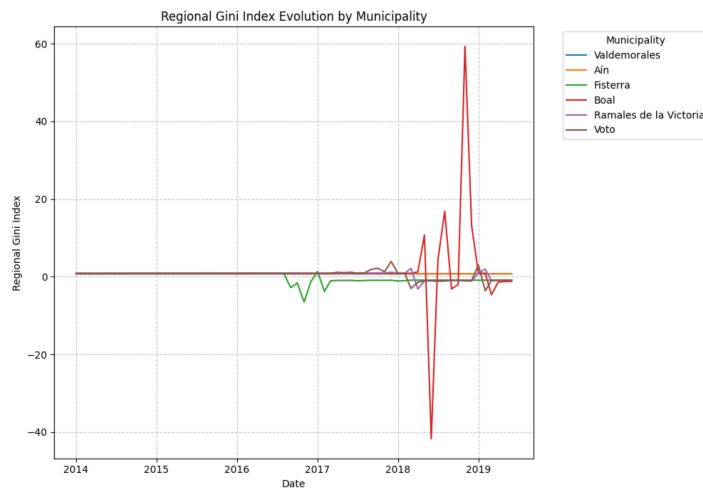


Figura 48: Evolución del índice de Gini regional en municipios seleccionados. Elaboración del autor.

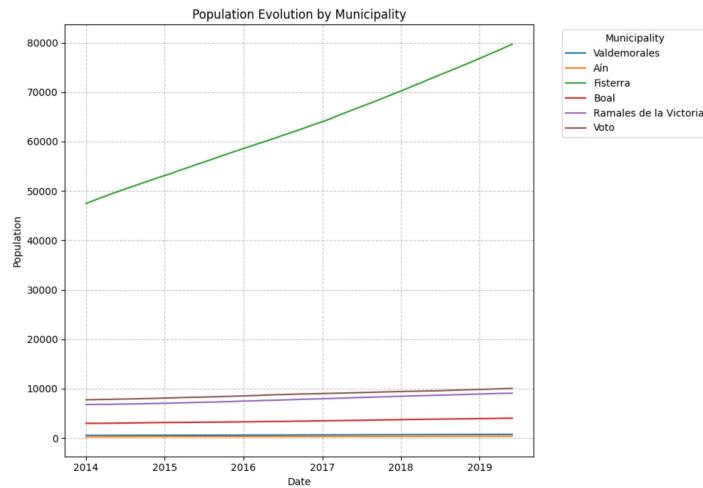


Figura 49: Evolución de la población en municipios seleccionados. Elaboración del autor.

Las comparativas puntuales en el último periodo de simulación (Figuras 50, 51, y 52) revelaron datos sobre permisos de construcción, tesorería municipal, y relaciones entre población, PIB, QLI y valor de la vivienda.

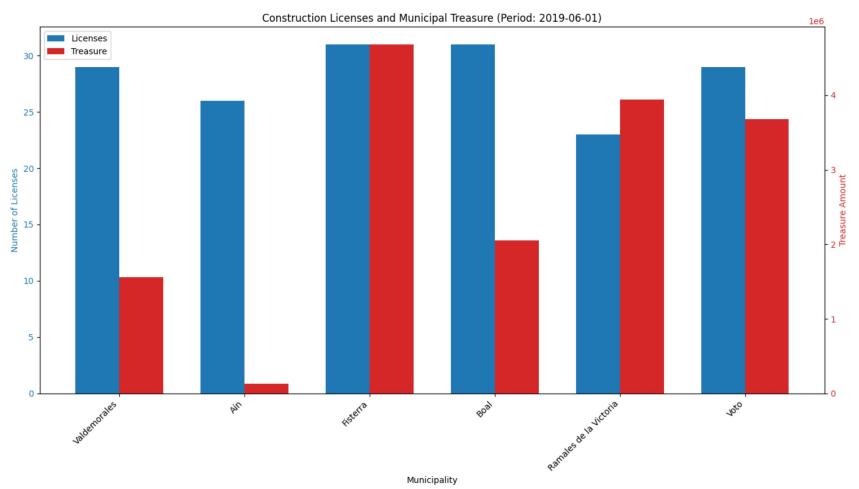


Figura 50: Comparativa de licencias de construcción y tesorería municipal (último periodo). Elaboración del autor.

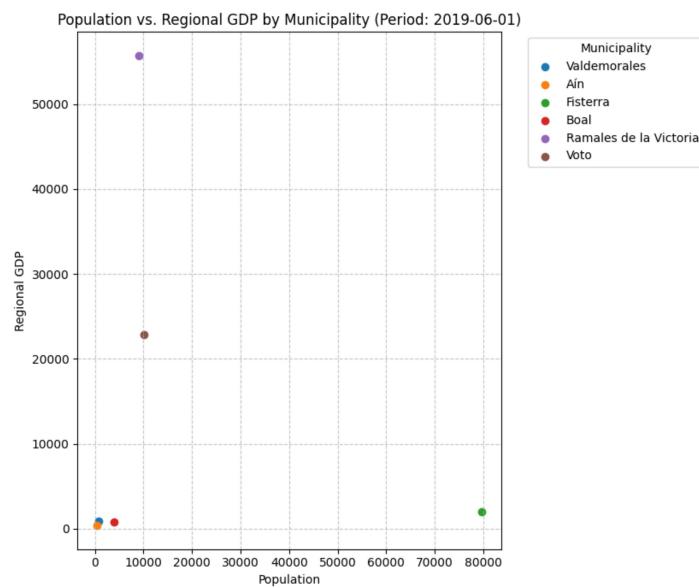


Figura 51: Relación entre población y PIB municipal (último periodo). Elaboración del autor.

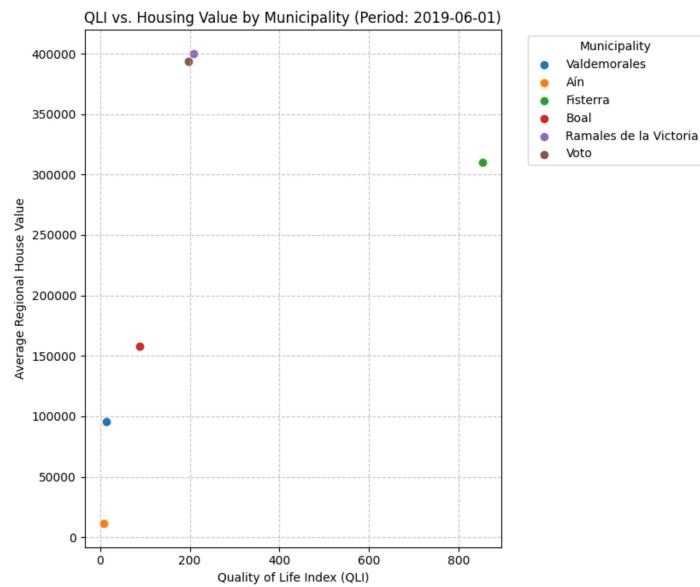


Figura 52: Relación entre QLI y valor de la vivienda municipal (último periodo). Elaboración del autor.

Los mapas coropléticos (Figuras 53 y 54) intentaron mostrar la distribución espacial del precio de venta y el QLI, aunque la infografía original presenta aún limitaciones en su visualización.

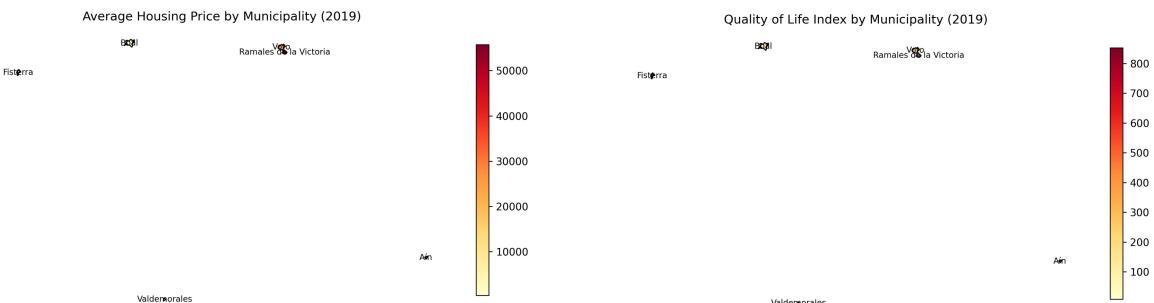


Figura 53: Mapa coroplético del precio de venta (simulado). Elaboración del autor.

Figura 54: Mapa coroplético del QLI (simulado). Elaboración del autor.

10.3.2. Impacto de Políticas Socioeconómicas Simuladas

La infografía comparativa se centró principalmente en el impacto de una "Política Salarial" (Wage Policy) con un coeficiente de intervención (**POLICY_COEFFICIENT**) de 0.8, frente al escenario "Sin Política". Los resultados para los municipios seleccionados indicaron que esta política salarial tenía la *mejora* (reducción) en los precios medios del alquiler (Figura 55) en comparación con el escenario base. Este efecto sugiere que, al menos bajo los parámetros simulados, un incremento salarial podría, contraintuitivamente, aliviar la presión sobre los alquileres, posiblemente a través de mecanismos indirectos como cambios en la demanda o la oferta de vivienda que el modelo captura. La Figura 57 muestra el impacto en los precios de

venta, y la Figura 56 en la tasa de desempleo, donde también se observan divergencias entre los escenarios. Otras comparativas regionales se muestran en la Figura 62.

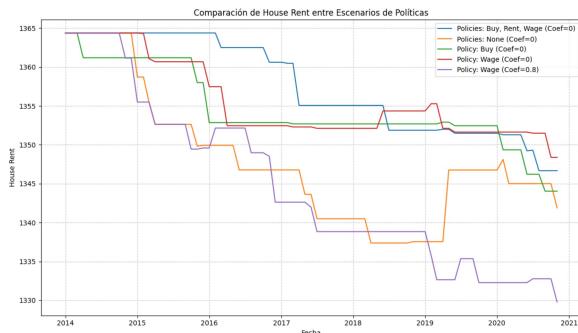


Figura 55: Impacto de la política salarial en el precio del alquiler. Elaboración del autor.



Figura 56: Impacto de la política salarial en la tasa de desempleo. Elaboración del autor.

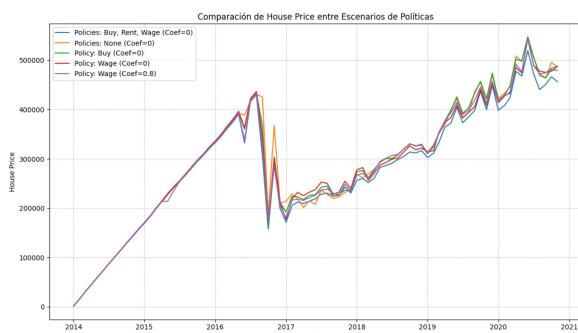


Figura 57: Impacto de la política salarial en el precio de venta de viviendas. Elaboración del autor.

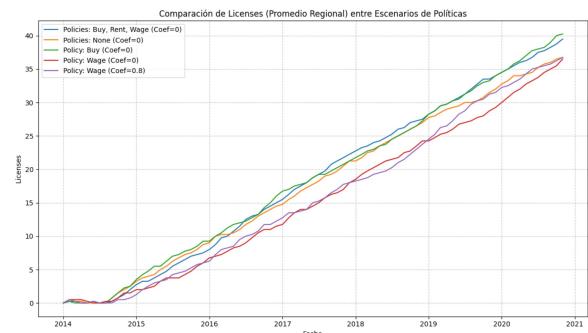


Figura 58: Impacto de la política salarial en licencias (promedio regional). Elaboración del autor.

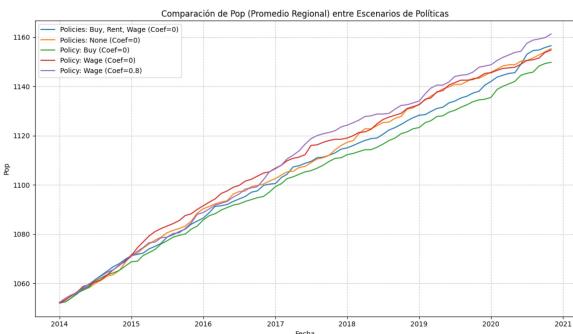


Figura 59: Impacto de la política salarial en población (promedio regional). Elaboración del autor.

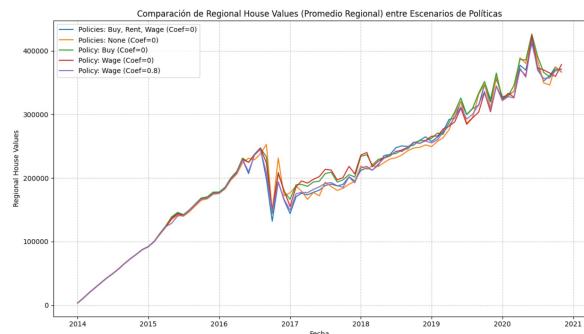


Figura 60: Impacto de la política salarial en valor de vivienda (promedio regional). Elaboración del autor.

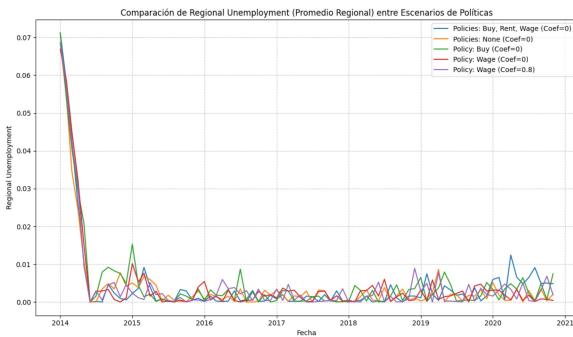


Figura 61: Impacto de la política salarial en desempleo (promedio regional). Elaboración del autor.

Figura 62: Impacto regional de la política salarial en población, valor de vivienda y desempleo. Elaboración del autor.

Es importante notar que la infografía se enfoca en esta política específica y no presenta una comparativa exhaustiva con otras políticas como los "Vales de Alquiler" para los mismos indicadores y municipios. La mención previa a dificultades con el parámetro **POLICY_COEFFICIENT=0** se refería a la visualización o análisis de ciertos escenarios si dicho coeficiente no se ajustaba adecuadamente, pero una vez se solucionaron problemas en la codificación se consiguió representar las políticas de manera adecuada como vemos para la política salarial con coeficiente 0.8, mostrando para el caso de estudio una reducción en los precios de alquiler, pareciendo esta medida más efectiva entre 2019-2021 y con baja repercusión en años anteriores.

10.3.3. Discusión sobre la Variabilidad y Calibración del Modelo

Los resultados de las simulaciones, incluso en el escenario base, muestran una variabilidad inherente a los modelos basados en agentes, donde las interacciones complejas pueden generar múltiples trayectorias. La calibración del modelo con datos reales españoles fue un paso fundamental para asegurar que estas dinámicas se mantuvieran dentro de rangos plausibles. La sensibilidad del modelo a parámetros como el **POLICY_COEFFICIENT** (observada en el caso de la política salarial) resalta la importancia de una validación exhaustiva y un análisis de sensibilidad de los componentes del modelo, especialmente aquellos que implementan las intervenciones de políticas. Para generalizar las conclusiones sobre el impacto de diversas intervenciones, aún es necesario explorar un espectro más amplio de escenarios de políticas y municipios, diferentes intensidades de aplicación (variando coeficientes como el **POLICY_COEFFICIENT**) y, potencialmente, realizar múltiples ejecuciones estocásticas para cada escenario para capturar el rango de posibles resultados.

10.4. Contribuciones a la Infraestructura de Datos

Una de las contribuciones fundamentales de este proyecto ha sido la creación y consolidación de una infraestructura de datos robusta y adaptada al contexto español, esencial para la simulación y el análisis del mercado inmobiliario. Esta infraestructura se materializa principalmente en una base de datos mejorada y en la integración de fuentes de datos detalladas como el Catastro.

10.4.1. Consolidación de la Base de Datos Mejorada

El proyecto PolicySpace2_Spain_new_ETL culminó con el desarrollo de una base de datos mejorada (alojada en `base_datos_mejorada/base_datos_mejorada.db`), que representa un avance significativo respecto a la gestión de datos fragmentados. Como se describe en su documentación (`README_BASE_DATOS.md` y `esquema_propuesto.md`), esta base de datos centraliza y estructura la información procesada a través de los diversos notebooks ETL. La consolidación incluye:

- **Tablas dimensionales geográficas:** Definen la jerarquía territorial (CCAA, provincias, municipios) y sus geometrías, permitiendo análisis espaciales precisos.
- **Tablas de hechos:** Contienen los datos procesados para variables clave como población, IDH y sus componentes, Participación en los Ingresos del Estado (PIE), número de empresas, tasas de mortalidad y fecundidad, tipos de interés, nivel educativo, tamaño medio de los hogares y, crucialmente, datos de viviendas catastrales.
- **Dataset integrado:** La creación del `common_dataset/dataset_municipio_cnae_anual_2014_2020.csv` es un ejemplo de producto de alto valor generado a partir de esta base de datos, ofreciendo una vista multidimensional a nivel municipal para un periodo específico.

Esta base de datos no solo sirve como el insumo principal para las simulaciones del modelo PolicySpace2 adaptado, sino que también facilita análisis exploratorios y la validación de datos a través de herramientas como el dashboard de Streamlit desarrollado. La Figura 21 presenta el esquema relacional de esta base de datos, ilustrando la interconexión de las diversas fuentes de información.

10.4.2. Integración de Datos Catastrales y su Impacto Potencial

La integración de datos de la Dirección General del Catastro (subproyecto CATASTRO) constituye otra contribución vital a la infraestructura de datos. Como se detalla en la Sección 5.10.12 y se ilustra en la Figura 23, se procesaron datos de edificaciones para el periodo 2014-2020, obteniendo información detallada como referencia catastral, coordenadas, superficie y año de alta.

El impacto de esta integración es doble:

- **Mayor realismo en la generación de viviendas:** Permite al modelo PolicySpace2, a través del método `_create_houses_real` en `world/generator.py`, utilizar datos de viviendas reales (Escenario 1, Figura 22) en lugar de depender únicamente de la generación sintética. Esto mejora significativamente la fidelidad de la representación del stock de viviendas y su localización espacial.
- **Base para análisis más detallados:** Aunque no explotado en toda su extensión en la fase actual, este dataset de viviendas catastrales abre la puerta a futuros análisis sobre la dinámica de la construcción, la antigüedad del parque inmobiliario, y la relación entre características físicas de las viviendas y su valor o uso.

La capacidad del modelo de recurrir a una generación sintética espacialmente consciente (Escenario 2, utilizando triangulación de polígonos municipales como se muestra en las Figuras 24, 25 y 26) cuando los datos catastrales no están disponibles, asegura la operatividad del modelo en todo el territorio nacional, pero la disponibilidad de datos catastrales reales siempre será preferible para estudios de mayor precisión.

Estas contribuciones a la infraestructura de datos no solo han sido cruciales para los análisis y simulaciones presentados en este trabajo, sino que también establecen una base sólida para futuras investigaciones y extensiones del modelo.

Recursos Digitales y Demostraciones del Proyecto

Esta sección recopila los principales recursos digitales, repositorios de código y demostraciones interactivas desarrolladas o utilizadas significativamente en el presente proyecto.

Nota importante sobre el acceso a los repositorios: En el momento de la redacción de este documento (junio de 2025), algunos de los repositorios de GitHub enlazados, particularmente aquellos que procesan datos catastrales originales o datos personales, se encuentran temporalmente en estado privado. Esto se debe a la necesidad de asegurar el cumplimiento de las licencias de uso de datos (ej. Catastro, que requiere transformación previa a la difusión) y la normativa de protección de datos. Se prevé la apertura pública de dichos repositorios una vez se hayan implementado las transformaciones, agregaciones o anonimizaciones pertinentes.

Cuadro 6: Principales Recursos Digitales del Proyecto.

Nombre del Recurso	Descripción Breve	Tipo	Enlace
Modelo Predictivo de Precios de Vivienda (proyecto_modelo_1)	Análisis y predicción de precios del mercado inmobiliario español usando Machine Learning. Implementación de varios modelos y análisis de importancia de variables.	Repositorio 	Visitar
Demostración del Modelo Predictivo	Sitio web interactivo que presenta los resultados, análisis de variables y visualizaciones del modelo predictivo de precios desarrollado.	Demo Web	Acceder
Datos Españoles para PolicySpace2 (PolicySpace2_Spanish_data)	Recopilación, procesamiento (ETL) y consolidación de datos demográficos, económicos y geográficos de España para alimentar el modelo PolicySpace2.	Repositorio 	Visitar
Dashboard de Datos (PolicySpace2_Spanish_data)	Aplicación Streamlit para la visualización y exploración de la base de datos (<code>datawarehouse.db</code>) y los resultados de los ETLs del proyecto PolicySpace2_Spanish_data.	Dashboard 	Acceder
Demo Web - Equivalencias PolicySpace2	Guía interactiva sobre equivalencias de datos España-Brasil para PolicySpace2 y acceso a recursos.	Demo Web	Acceder
Demo Web - Adaptación PolicySpace2	Presentación de la adaptación de PolicySpace2 al contexto español, fuentes de datos y scripts.	Demo Web	Acceder
Adaptación de PolicySpace2 a España (PolicySpace2_Spain_new_ETL)	Refinamiento del proceso ETL y creación de una base de datos mejorada para la adaptación funcional del modelo PolicySpace2 al contexto español.	Repositorio 	Visitar
Integración de Datos Catastrales (CATASTRO)	Procesamiento de datos de la Dirección General del Catastro para la generación realista de viviendas y su integración en el modelo de simulación.	Repositorio 	Visitar

11. Conclusiones y trabajos futuros

La adaptación del modelo *PolicySpace2* al contexto municipal español ha permitido integrar datos reales de demografía, economía y vivienda en la simulación. Gracias a la incorporación de variables municipales auténticas (población, viviendas catastrales, indicadores socioeconómicos, etc.), el modelo reproduce dinámicas locales más realistas. En particular, se han cargado tablas geográficas y demográficas detalladas (comunidades autónomas, provincias, municipios, población, mortalidad, IDH, fecundidad, vivienda), lo que sustenta análisis precisos de políticas públicas a nivel local.

Como trabajos futuros, se plantean las siguientes líneas específicas:

- **Solución de bugs en el índice de Gini:** Continuar corrigiendo errores en el cálculo del índice de Gini regional. Por ejemplo, en el análisis post-simulación se identificó un filtrado temporal incorrecto (usando diciembre de 2019 en lugar de 2014) que impedía obtener valores de Gini. Esta incidencia fue corregida ajustando el filtro al mes correcto (no coincidente con la gráfica mostrada en el informe), lo que permitió generar resultados válidos de Gini regional. Se continuará validando este indicador y solucionando posibles discrepancias con datos empíricos externos.
- **Inclusión de datos censales internos:** Incorporar microdatos censales y estadísticos municipales (por ejemplo, datos de sección censal del INE) para capturar dinámicas demográficas y sectoriales internas. Esto permitirá analizar fenómenos específicos del mercado de vivienda, como el auge de los alquileres vacacionales (Airbnb), que ha elevado los precios residenciales y reducido la oferta de vivienda para residentes locales. Datos censales detallados facilitarán estudiar la movilidad interna y la segregación residencial, mejorando la modelización de la demanda de vivienda, así como la posibilidad de poder aplicar esta herramienta para comparar resultados según política aplicada en cada municipio o incluso la posibilidad de tener un seguimiento de políticas aplicadas con un gran detalle en resultados tanto sociales como económicos para medir sus efectos.
- **Integración del modelo predictivo de precios de vivienda:** Incorporar el modelo predictivo desarrollado en el repositorio `proyecto_modelo_1`. Este modelo emplea técnicas de aprendizaje automático (por ejemplo, regresión multivariante) para estimar el valor de las viviendas en función de características estructurales (área, calidad, ubicación, etc.). Su lógica, descrita en el informe proporcionado, se utilizará para refinar la fórmula base de cálculo de precios. La fórmula actual del modelo define el precio inicial como

$$P = \text{ÁREA} \times \text{precio}_{m^2}$$

, con $\text{precio}_{m^2} = 1000$ ($\text{€}/m^2$ fijo). Con el modelo predictivo, la estimación del precio sería del tipo

$$P = \alpha + \beta \text{ÁREA} + \gamma \text{Calidad} + \delta \text{Ubicación} + \dots,$$

ajustando los coeficientes $\alpha, \beta, \gamma, \delta, \dots$ a los datos observados. Esto mejoraría la precisión y adaptabilidad del precio simulado al contexto real.

- **Generación de escenarios demográficos con IA:** Desarrollar y aplicar modelos de proyección basados en inteligencia artificial (por ejemplo, redes neuronales generativas o modelos de series temporales avanzados) para simular variables demográficas clave (natalidad, mortalidad, migración). Estos escenarios futuros permitirán alimentar la simulación con trayectorias plausibles a largo plazo, mejorando la planificación de políticas municipales y la evaluación de tendencias sociales.
- **Evaluación de políticas fiscales diferenciadas por municipio:** Analizar el impacto de políticas fiscales locales (variaciones en impuestos municipales, subvenciones, etc.) en los indicadores económicos y sociales del modelo. Esto permitiría simular cómo distintas estrategias impositivas afectan la recaudación local, la inversión pública y la equidad intermunicipal, ayudando a diseñar políticas fiscales más efectivas.
- **Movilidad diaria (commuting) intermunicipal:** Incorporar datos reales de flujos de movilidad laboral y educativa entre municipios. Modelar la movilidad cotidiana de habitantes (trabajo, estudio) enriquecerá el contexto espacial del modelo, permitiendo capturar la dependencia económica entre municipios y su efecto en la demanda de vivienda. Por ejemplo, el commuting influye en la distribución de ingresos familiares y en la presión sobre mercados de vivienda en zonas dormitorio o centros urbanos.

A continuación se presenta una tabla comparativa que resume la lógica de determinación de precios de vivienda en el modelo original y en la propuesta que integra el modelo predictivo. En el modelo actual, el precio se calcula de forma lineal como $P = \text{ÁREA} \times 1000$ ($\text{€}/\text{m}^2$ fijo). Con el modelo predictivo, el precio se estima mediante una regresión ajustada a datos reales, incorporando múltiples variables (área, calidad, ubicación, etc.). La Tabla 7 resume estas diferencias clave:

Aspecto	Modelo actual	Con modelo predictivo
Fórmula de cálculo	$P = \text{ÁREA} \times 1000$ (precio m^2 fijo)	$P = \alpha + \beta \text{ÁREA} + \gamma \text{Calidad} + \dots$
Variables consideradas	Solo la superficie de la vivienda y parámetro fijo de precio	Múltiples: superficie, calidad, ubicación, características del entorno, etc.
Dependencia espacial	No distingue entre barrios (precio homogéneo por municipio)	Ajusta el precio según datos locales y atributos observados en cada municipio
Precisión y realismo	Limitado (no capta la heterogeneidad real del mercado)	Superior (estima precios basados en datos históricos y atributos reales)

Cuadro 7: Comparación de la lógica de precios actual del modelo y la propuesta con el modelo predictivo.

Glosario

Inteligencia Artificial (IA) Campo de la informática que se centra en el desarrollo de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, la toma de decisiones y el reconocimiento de patrones.

Big Data Conjunto de datos extremadamente grandes y complejos que requieren herramientas avanzadas para su almacenamiento, procesamiento y análisis.

Gradient Boosting Machine (GBM) Modelo de machine learning basado en árboles de decisión, que optimiza iterativamente los errores residuales para mejorar la precisión predictiva.

Random Forest (RF) Conjunto de árboles de decisión que trabajan juntos para mejorar la precisión y reducir el riesgo de sobreajuste en predicciones.

ARIMA Modelo estadístico utilizado en análisis de series temporales para entender y predecir patrones en datos históricos.

Prophet Modelo de análisis de series temporales desarrollado por Facebook, diseñado para detectar patrones estacionales y tendencias.

Clustering Método de agrupamiento que organiza datos en grupos homogéneos basados en características similares.

Gentrificación Proceso de transformación urbana que desplaza a los residentes originales debido al aumento de precios y la llegada de nuevos habitantes con mayor poder adquisitivo.

Econometría Rama de la economía que utiliza métodos estadísticos y matemáticos para analizar y evaluar relaciones económicas.

ETL (Extract, Transform, Load) Proceso de extracción, transformación y carga de datos desde múltiples fuentes hacia un almacenamiento unificado para su análisis.

Reglamento General de Protección de Datos (RGPD) Normativa europea que regula el tratamiento de datos personales, garantizando derechos de privacidad y seguridad para los ciudadanos.

AWS (Amazon Web Services) Plataforma de servicios en la nube que ofrece almacenamiento, procesamiento y otras soluciones tecnológicas escalables.

LLM (Large Language Models) Modelos de lenguaje a gran escala entrenados con vastos volúmenes de datos para comprender, generar y manipular texto de manera coherente y contextual.

Machine Learning (ML) Rama de la inteligencia artificial que utiliza algoritmos y modelos estadísticos para que los sistemas informáticos puedan aprender y mejorar automáticamente a partir de los datos sin ser programados explícitamente.

Web Scraping Técnica automatizada para extraer información de sitios web en formato estructurado para su posterior análisis.

Script Conjunto de instrucciones o código que se utiliza para automatizar tareas específicas en un programa o sistema.

Redes Neuronales Modelos computacionales inspirados en la estructura y el funcionamiento del cerebro humano, utilizados para identificar patrones y realizar predicciones en el ámbito del aprendizaje automático.

Agentes Entidades autónomas utilizadas en modelos computacionales para simular el comportamiento de actores individuales o colectivos en un entorno definido, como compradores, vendedores o reguladores. Son esenciales en enfoques de Modelos Basados en Agentes (ABM, por sus siglas en inglés).

ABM Modelos Basados en Agentes (por sus siglas en inglés, Agent-Based Models), una técnica de simulación computacional que utiliza agentes autónomos para modelar y analizar dinámicas complejas en sistemas sociales, económicos o ambientales.

RMSE Raíz del error cuadrático medio (Root Mean Squared Error, por sus siglas en inglés), una métrica que evalúa la precisión de los modelos predictivos midiendo el promedio de las diferencias cuadradas entre los valores observados y los valores predichos.

R² Coeficiente de determinación, una métrica que mide la proporción de la varianza en la variable dependiente que puede ser explicada por las variables independientes en un modelo.

Índice de Gini Medida estadística de la desigualdad en una distribución, comúnmente utilizada para analizar la desigualdad económica. Un valor de 0 indica igualdad perfecta y un valor de 1 representa desigualdad máxima.

Proxies

Ciencia de Datos (DS) Disciplina interdisciplinaria que combina estadística, informática y conocimiento del dominio para recopilar, procesar, analizar y visualizar datos, con el fin de extraer información valiosa y apoyar la toma de decisiones basadas en evidencia.

Referencias

- [1] Alejandro Baldominos, Iván Blanco, Antonio José Moreno, Rubén Iturrarte, Óscar Bernárdez, and Carlos Afonso. Identifying real estate opportunities using machine learning. *Applied Sciences (Switzerland)*, 8, 2018. URL: <https://arxiv.org/pdf/1809.04933.pdf>, doi:10.3390/app8112321.
- [2] Agustin Cocola-Gant and Antonio Lopez-Gay. Transnational gentrification, tourism and the formation of ‘foreign only’ enclaves in barcelona. *Urban Studies*, 57, 2020. URL: <https://journals.sagepub.com/doi/epub/10.1177/0042098020916111>, doi:10.1177/0042098020916111.
- [3] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9, 2017. URL: <https://www.aeaweb.org/articles?id=10.1257/app.20160213>, doi:10.1257/app.20160213.
- [4] Consejo Europeo. Reglamento general de protección de datos, 2024. URL: <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679>.
- [5] Bernardo Alves Furtado. PolicySpace2: Modeling markets and endogenous public policies. *JASSS*, 25, 2022. doi:10.18564/jasss.4742.
- [6] Winky K.O. Ho, Bo Sin Tang, and Siu Wai Wong. Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38, 2021. doi:10.1080/09599916.2020.1832558.
- [7] Fabiënne Kortas, Alexander Grigoriev, and Giulia Picollo. Exploring multi-scale variability in hotspot mapping: A case study on housing prices and crime occurrences in heerlen. *Cities*, 128:103814, 2022. doi:10.1016/j.cities.2022.103814.
- [8] Jaehun Lee, Taewon Suh, Daniel Roy, and Melissa Baucus. Emerging technology and business model innovation: The case of artificial intelligence. *Journal of Open Innovation: Technology, Market, and Complexity*, 5, 2019. URL: <https://www.sciencedirect.com/science/article/pii/S2199853122009817>, doi:10.3390/joitmc5030044.
- [9] José Miguel Hernández López. *Reglamento de Inteligencia Artificial. Incluye introducción, notas, cronología, webgrafía, bibliografía e índice analítico*. 2024. URL: [https://app.vlex.com/search/jurisdiction:ES,EU/El+reglamento+\(UE\)+INTELIGENCIA+ARTIFICIAL/vid/1048850667](https://app.vlex.com/search/jurisdiction:ES,EU/El+reglamento+(UE)+INTELIGENCIA+ARTIFICIAL/vid/1048850667).
- [10] Shirley Nieuwland and Rianne van Melik. Regulating airbnb: how cities deal with perceived negative externalities of short-term rentals. *Current Issues in Tourism*, 23, 2020. URL: https://repub.eur.nl/pub/109968/REPUB_109968_OA.pdf, doi:10.1080/13683500.2018.1504899.
- [11] Roberto Verganti, Luca Vendraminelli, and Marco Iansiti. Innovation and design in the age of artificial intelligence. *Journal of Product Innovation Management*, 37, 2020. URL: https://www.hbs.edu/ris/Publication%20Files/20-091_3889aa72-1853-42f8-8b17-5760c86f863e.pdf, doi:10.1111/jpim.12523.
- [12] Ismael Yrigoy. Airbnb en menorca: ¿una nueva forma de gentrificación turística?: Localización de la vivienda turística, agentes e impactos sobre el alquiler residencial. *Scripta Nova*, 21, 2017. URL: <https://revistes.ub.edu/index.php/ScriptaNova/article/view/18573/22698>, doi:10.1344/sn2017.21.18573.
- [13] Miquel Àngel Garcia-López, Jordi Jofre-Monseny, Rodrigo Martínez-Mazza, and Mariona Segú. Do short-term rental platforms affect housing markets? evidence from airbnb in barcelona. *Journal of Urban Economics*, 119, 2020. URL: <https://www.sciencedirect.com/science/article/pii/S0094119020300498>, doi:10.1016/j.jue.2020.103278.