Proyecto: Análisis y predicción basados en datos de precios de vivienda en el estado de Washington, Estados Unidos

Grupo 2: Lizeth García, Leonardo Guzmán, María Alejandra Herrera, Carlos Silva

1. Contexto

El mercado inmobiliario estadounidense se caracteriza por una alta diversidad en los precios de las propiedades, determinada por una combinación de factores estructurales (como el tamaño, número de habitaciones, antigüedad y materiales de construcción) y factores externos (ubicación geográfica, escasez de oferta, incremento en las tasas hipotecarias y cambios en las dinámicas de consumo tras la pandemia). Esta complejidad genera dificultades para estimar con precisión el valor real de una vivienda, lo que puede derivar en sobrevaloraciones o subvaloraciones que afectan la eficiencia en la toma de decisiones de compra, inversión y financiamiento.

La *Federal Housing Finance Agency* (FHFA), entidad reguladora del sistema hipotecario estadounidense, ha reportado que el Índice de Precios de Viviendas (*House Price Index* - HPI) alcanzó su máximo histórico en febrero de 2025. Aunque el ritmo de crecimiento se ha moderado recientemente, la tendencia de valorización continúa impactando la accesibilidad al mercado y la competitividad del sector inmobiliario.

Ante este panorama, el presente proyecto propone el desarrollo de un modelo de analítica predictiva basado en técnicas de aprendizaje supervisado, orientado a estimar el precio de venta de una vivienda a partir de sus características físicas y contextuales. Para ello, se utilizará un conjunto de datos históricos disponible en Kaggle, que requiere de procesos de limpieza, transformación y análisis exploratorio para garantizar su calidad y utilidad analítica.

En conclusión, este proyecto busca demostrar cómo la analítica predictiva puede convertirse en un recurso estratégico para mejorar la eficiencia en sectores tradicionales, transformando información histórica en conocimiento útil y aplicable, y contribuyendo a la transparencia y racionalidad del mercado inmobiliario.

2. Pregunta de negocio

¿Cuáles son los factores estructurales y geográficos que determinan el precio de una vivienda en el estado de Washington en Estados Unidos, y cómo pueden utilizarse para estimar su valor esperado antes de salir al mercado, con el fin de apoyar decisiones estratégicas de inversión, comercialización y financiamiento en el sector inmobiliario?

Esta pregunta busca identificar las variables más influyentes en la formación del precio de venta de una propiedad, considerando tanto sus características físicas y constructivas (como tamaño, número de habitaciones, antigüedad o materiales) como su entorno geográfico y económico (ubicación, vecindario, oferta y demanda regional).

3. Alcance del proyecto

El proyecto contempla el desarrollo de un modelo supervisado de aprendizaje automático, basado en datos históricos del mercado inmobiliario estadounidense disponibles en Kaggle. Dicho modelo permitirá estimar el valor esperado de una vivienda antes de su comercialización, a partir de las variables más relevantes identificadas en el análisis exploratorio.

El alcance incluye:

- La limpieza, transformación y análisis exploratorio del conjunto de datos.
- La construcción, entrenamiento y validación de modelos predictivos.
- El despliegue del modelo a través de una API interactiva, integrada con un tablero de visualización que facilite la interpretación de los resultados y el uso práctico del modelo por parte de los usuarios.

La herramienta busca facilitar decisiones informadas y basadas en evidencia dentro del sector inmobiliario, ofreciendo una solución ágil, interactiva y respaldada por datos. Podría ser útil para compradores, inversionistas, agentes inmobiliarios y entidades financieras, al reducir el riesgo de sobrevaloración o subvaloración, anticipar tendencias del mercado y mejorar la eficiencia en la toma de decisiones.

4. Descripción de conjuntos de datos a emplear

El conjunto de datos denominado *USA House Prices*, disponible en Kaggle[1], contiene información detallada sobre propiedades residenciales vendidas en el estado de Washington en Estados Unidos, recopilada con el objetivo de facilitar el análisis y modelado de precios de venta de viviendas.

Dentro de las principales características del conjunto de datos, están:

- Formato: CSV
- Tamaño: Aproximadamente 1 MB
- Número de registros: 4140 que corresponden a propiedad
- Variables incluidas (definidas en Kaggle):
    - Date: La fecha en la que se dio la venta de la vivienda.

[1] Disponible en línea en https://www.kaggle.com/datasets/fratzcan/usa-house-prices. Octubre 2025.

- Price: El precio al que se vendió la vivienda en USD.
- Bedrooms: Cantidad de cuartos de la vivienda.
- Bathrooms: Cantidad de baños de la vivienda.
- Sqft Living: El área de la sala en pies cuadrados.
- Sqft Lot: El área del lote de la vivienda.
- Floors: La cantidad de pisos de la vivienda.
- Waterfront: Variable binaria que indica si la Vivienda tiene vista al agua o no.
- wise).
- View: Indice de 0 a 4 que indica la calidad de la vista de la vivienda
- Condition: Indice de 1 a 5 que indica la condición de la vivienda.
- Sqft Above: Área de la vivienda sin contar el sótano en pies cuadrados.
- Sqft Basement: Área del sótano de la vivienda en pies cuadrados.
- Yr Built: Año en el que se construyó la vivienda.
- Yr Renovated: El año en el que se remodeló la vivienda.
- Street: La dirección de la vivienda.
- City: La ciudad donde se encuentra la vivienda.
- Statezip: El estado y código postal de la vivienda.
- Country: El país donde se encuentra la vivienda.

5. Exploración de datos

Para la exploración de los datos se creo un notebook en Python para cargar la información en un Dataframe de pandas y proceder con los análisis mostrados a continuación. Para mayor detalle, el notebook se encuentra alojado en el repositorio remoto de GitHub: https://github.com/carlossil05/G2-Proyecto-DSA.

Para la exploración inicial se utilizan las funciones (head(), describe(), groupby()) y atributos (shape) de los dataframes de pandas. Se evidencia lo siguiente:

- Como se puede ver en la Figura 1, el conjunto de datos cuenta con un total de 4140 observaciones y 18 variables, de las cuales hay 4 variables categóricas, una variable de fechas y 13 variables numéricas.

```
#Cargar datos a un dataframe
df=pd.read_csv('./data/USAHousingDataset.csv')
#Cantidad de observaciones y variables del dataset
df.shape
```

```
(4140, 18)
```

```
#Tipo de datos del dataset
df.dtypes
```

```
date              object
price            float64
bedrooms         float64
bathrooms        float64
sqft_living        int64
sqft_lot           int64
floors           float64
waterfront         int64
view               int64
condition          int64
sqft_above         int64
sqft_basement      int64
yr_built           int64
yr_renovated       int64
street            object
city              object
statezip          object
country           object
dtype: object
```

Figura 1. Cantidad de datos y tipo de datos del conjunto.

- Las primeras 5 observaciones del conjunto de datos se pueden observar en la Figura 2, evidenciando variables de precio, características de la vivienda como numero de cuartos y baños, el área y los pisos, la vista de la vivienda, los años de construcción y renovación y la ubicación de cada vivienda que incluye calle, ciudad, estado, código postal y país.

```
#Vista de las primeras observaciones del dataset
df.head()
```

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
0	2014-05-09 00:00:00	376000.0	3.0	2.00	1340	1384	3.0	0	0
1	2014-05-09 00:00:00	800000.0	4.0	3.25	3540	159430	2.0	0	0
2	2014-05-09 00:00:00	2238888.0	5.0	6.50	7270	130017	2.0	0	0
3	2014-05-09 00:00:00	324000.0	3.0	2.25	998	904	2.0	0	0
4	2014-05-10 00:00:00	549900.0	5.0	2.75	3060	7015	1.0	0	0

condition	sqft_above	sqft_basement	yr_built	yr_renovated	street	city	statezip	country
3	1340	0	2008	0	9245-9249 Fremont Ave N	Seattle	WA 98103	USA
3	3540	0	2007	0	33001 NE 24th St	Carnation	WA 98014	USA
3	6420	850	2010	0	7070 270th Pl SE	Issaquah	WA 98029	USA
3	798	200	2007	0	820 NW 95th St	Seattle	WA 98117	USA
5	1600	1460	1979	0	10834 31st Ave SW	Seattle	WA 98146	USA

Figura 2. Vista de las primeras filas del conjunto de datos.

- De las estadísticas descriptivas de las variables numéricas se puede ver que:
    - El promedio del precio de la vivienda está alrededor de los $553,000USD. La vivienda más cara de la muestra se vendió en $26,590,000USD.
    - En promedio las viviendas tienen 3 cuartos y 2 baños, y entre 1 y 2 pisos.
    - La mayoría de las viviendas se construyeron en promedio en el año 1970 con una desviación estándar de 29.8 años.
    - El área de promedio del lote de las viviendas es de 14,697 pies cuadrados.

```
#Estadísticas descriptivas de las variables numéricas
df.describe()
```

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
count	4.140000e+03	4140.000000	4140.000000	4140.000000	4.140000e+03	4140.000000
mean	5.530629e+05	3.400483	2.163043	2143.638889	1.469764e+04	1.514130
std	5.836865e+05	0.903939	0.784733	957.481621	3.587684e+04	0.534941
min	0.000000e+00	0.000000	0.000000	370.000000	6.380000e+02	1.000000
25%	3.200000e+05	3.000000	1.750000	1470.000000	5.000000e+03	1.000000
50%	4.600000e+05	3.000000	2.250000	1980.000000	7.676000e+03	1.500000
75%	6.591250e+05	4.000000	2.500000	2620.000000	1.100000e+04	2.000000
max	2.659000e+07	8.000000	6.750000	10040.000000	1.074218e+06	3.500000

waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated
4140.000000	4140.000000	4140.000000	4140.000000	4140.000000	4140.000000	4140.000000
0.007488	0.246618	3.452415	1831.351449	312.287440	1970.814010	808.368357
0.086219	0.790619	0.678533	861.382947	464.349222	29.807941	979.380535
0.000000	0.000000	1.000000	370.000000	0.000000	1900.000000	0.000000
0.000000	0.000000	3.000000	1190.000000	0.000000	1951.000000	0.000000
0.000000	0.000000	3.000000	1600.000000	0.000000	1976.000000	0.000000
0.000000	0.000000	4.000000	2310.000000	602.500000	1997.000000	1999.000000
1.000000	4.000000	5.000000	8020.000000	4820.000000	2014.000000	2014.000000

Figura 3. Estadísticas descriptivas de las variables numéricas del conjunto de datos.

- De las estadísticas descriptivas de las variables categóricas se puede ver que:
    - El conjunto de datos contiene viviendas de 43 ciudades distintas de USA y distribuidos en un total de 77 códigos postales.
    - La mayoría de las observaciones se encuentran en la ciudad de Seattle.
    - El código postal más frecuente en el con junto de datos es 98103.
    - Todas las observaciones corresponden al verano del año 2014 entre los meses de mayo y julio.

```
#Estadística descriptiva de las variables categóricas
df.describe(include='object')
```

	date	street	city	statezip	country
count	4140	4140	4140	4140	4140
unique	68	4079	43	77	1
top	2014-06-23 00:00:00	2520 Mulberry Walk NE	Seattle	WA 98103	USA
freq	142	4	1415	128	4140

Figura 4. Estadísticas descriptivas de las variables categóricas del conjunto de datos.

```
#Convertir columna date a datetime para facilitar exploración
df['date'] = pd.to_datetime(df['date'])
#Se agrupan por mes para evaluar la fecha de las observaciones
df.groupby(df['date'].dt.to_period('M')).size()
```

```
date
2014-05    1308
2014-06    2179
2014-07     653
Freq: M, dtype: int64
```

Figura 5. Datos agrupados por mes.

Se construyeron histogramas para ilustrar y complementar el análisis sobre la distribución de los datos como se muestra en la Figura 6 para algunas de las variables. Se observa que la mayoría de las viviendas están concentradas en un rango de precios entre 0 y $1,000,000USD y también se observa una distribución del año de construcción de la vivienda que aumenta hacia el año 2000.

```python
#Histogramas de los datos
numeric_cols = df.select_dtypes(include='number').columns

#Para cada columna numérica se grafica un histograma
for col in numeric_cols:
    plt.figure(figsize=(6, 4))
    sns.histplot(df[col], bins=30)
    plt.title(f"Distribución de {col}")
    plt.xlabel(col)
    plt.ylabel("Cuenta")
    plt.show()
```

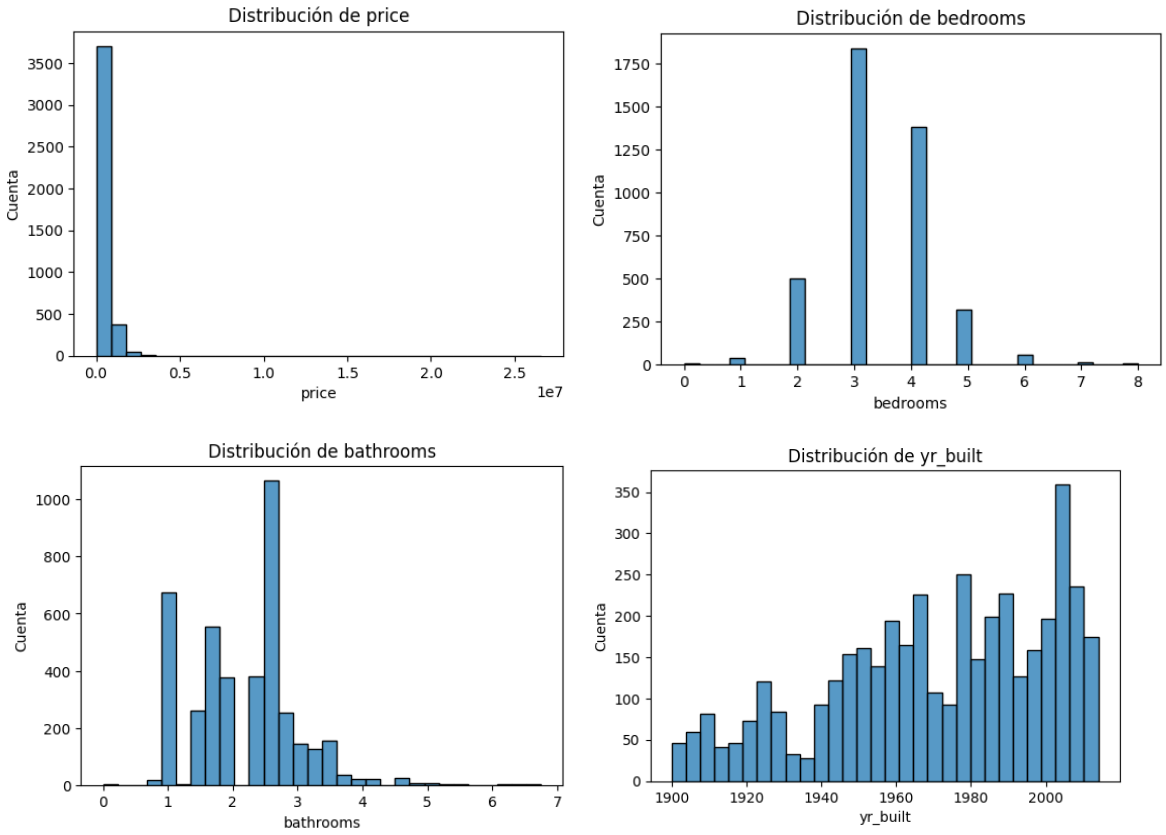

Figura 6. Histogramas de algunas de las variables del conjunto de datos.

Se construyó un mapa de color basado en la correlación entre las variables numéricas del conjunto de datos, ver Figura 7, para evaluar la influencia de los factores estructurales y características de las viviendas en su precio de venta y hacer un primer acercamiento sobre cuales de estas variables pueden ser las más influyentes para responder a la pregunta de negocio:

- Se observa que las variables numéricas con mayor correlación frente al precio son la cantidad de baños, el área de la sala y el área de la vivienda sin contar el sótano.
- Se observa que la correlación del precio de venta es prácticamente cero con las variables área del lote, condición de la vivienda y los años de construcción y renovación. Se debe revisar con mayor profundidad por qué estas variables no están influenciando el precio de venta.

```
plt.figure(figsize=(8, 6))
sns.heatmap(corr, annot=True, cmap='coolwarm', center=0)
plt.title("Mapa de calor de correlación de variables numéricas")
plt.show()
```

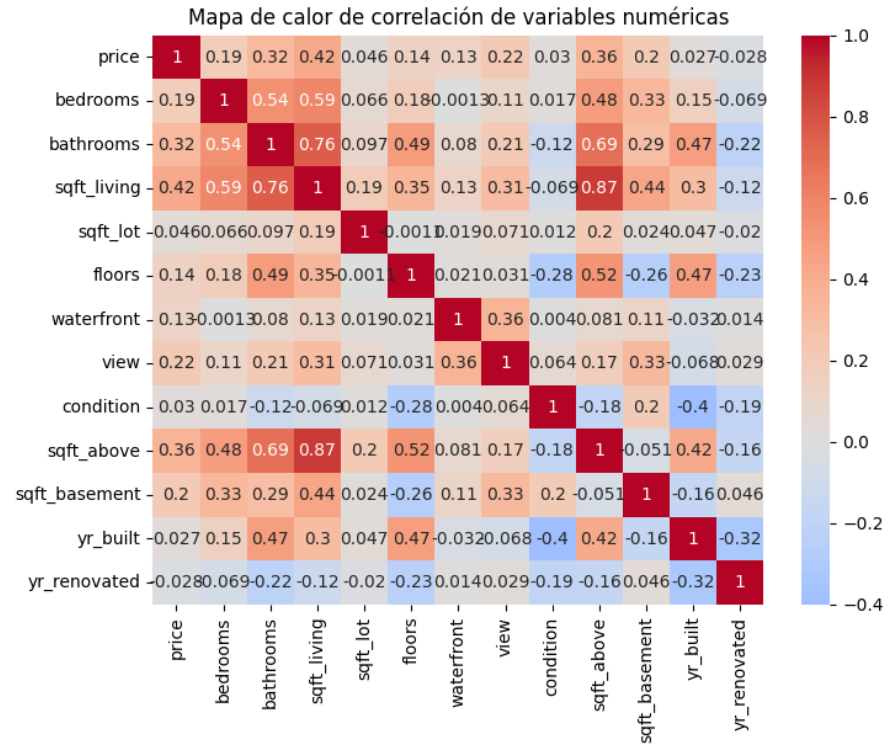

Figura 7. Mapa de calor de correlación entre las variables numéricas del conjunto de datos.

Finalmente, se construyó el mapa de calor geográfico de precio por área de las viviendas mostrado en la Figura 8 para dar una respuesta inicial en cuánto a que tanto puede influir la ubicación de la vivienda en el precio de venta. Se observa que el precio de la vivienda es mayor cuando se encuentra dentro de la ciudad de Seattle y que tiende a disminuir a medida que se aleja de su casco urbano.

```
[14]: precio_ciudad = df.groupby('city')['price_per_sqft'].mean().reset_index()

      # Inicializar geolocator
      geolocator = Nominatim(user_agent="wa_price_map")
      geocode = RateLimiter(geolocator.geocode, min_delay_seconds=1)

      # Se agrega el estado de Washington en USA a cada ciudad para facilitar el geocode
      def safe_geocode(city):
          try:
              return geocode(f"{city}, Washington, USA")
          except:
              return None

      #Encontrar las coordenadas geográficas de cada ciudad
      precio_ciudad['location'] = precio_ciudad['city'].apply(safe_geocode)
      precio_ciudad['lat'] = precio_ciudad['location'].apply(lambda loc: loc.latitude if loc else None)
      precio_ciudad['lon'] = precio_ciudad['location'].apply(lambda loc: loc.longitude if loc else None)
```

```
[15]: # Gráfica geográfica usando plotly express scatter_map
      fig = px.scatter_map(
          precio_ciudad,
          lat='lat',
          lon='lon',
          size='price_per_sqft',
          color='price_per_sqft',
          hover_name='city',
          color_continuous_scale='Viridis',
          map_style='carto-positron',
          zoom=7,
          title='Precio promedio por sqft por ciudad',
          width=700,
          height=700,
      )
      fig.show()
```

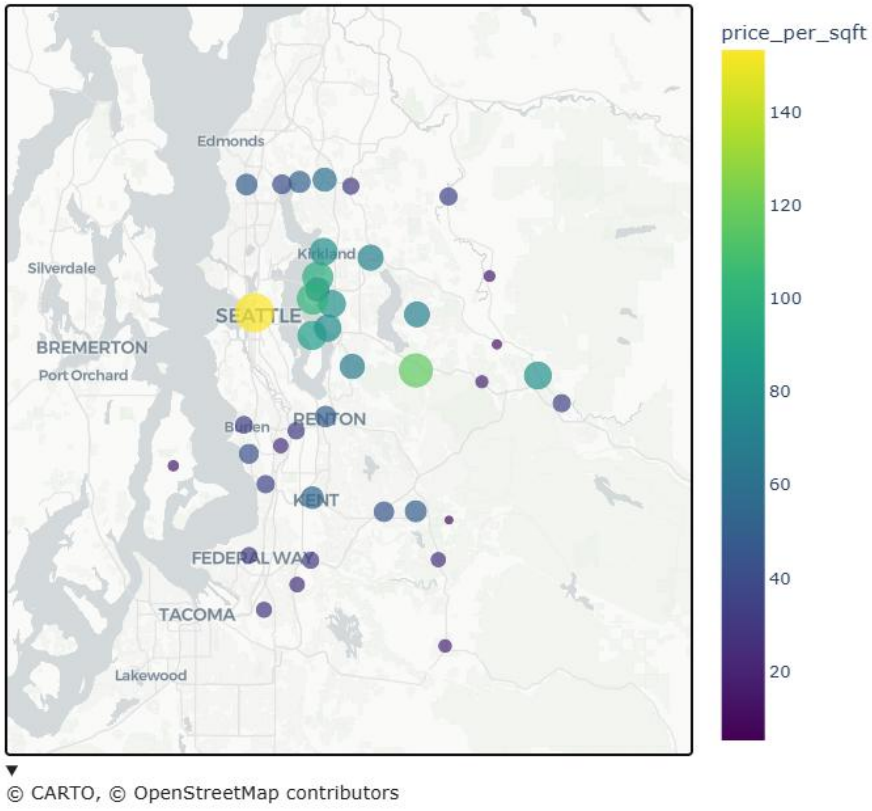

Figura 8. Mapa de calor del precio por área de las viviendas.

6. Maqueta del prototipo

Se propone un prototipo que simplifique el procedimiento de la predicción del precio de venta de la vivienda y que también muestre un mapa interactivo el comparativo de precios por zona o precios de viviendas similares a la que se quiere evaluar. El usuario únicamente deberá ingresar los datos que se soliciten en el formulario de las variables predictoras más significativas que se definan para el modelo.

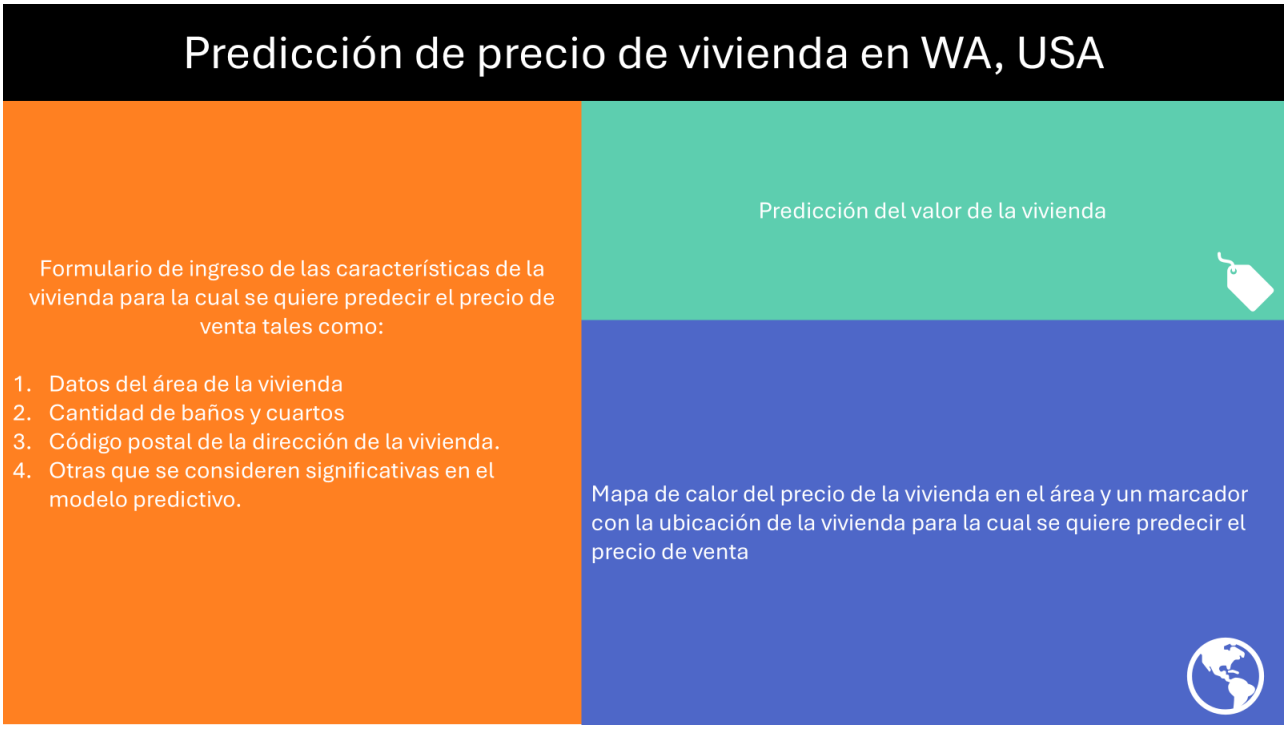

Figura 9. Maqueta del prototipo.

7. Reporte de trabajo en equipo

El equipo trabajó de manera colaborativa y coordinada, realizando reuniones periódicas para planificar avances, asignar tareas y revisar resultados. Cada una de las actividades se gestionaron mediante herramientas digitales que facilitaron la comunicación y el seguimiento del proyecto:

- GitHub: para el control de versiones del código y la integración de los aportes individuales.
- Google Drive: para compartir los avances y entregables
- Google Meet: para realizar reuniones de revisión, resolver dudas y validar resultados.

Estas herramientas permitieron mantener una comunicación constante, garantizar la trazabilidad del trabajo y asegurar la consistencia de las versiones del proyecto.

Cada integrante participó activamente en las fases desarrolladas hasta la fecha: definición del problema, análisis exploratorio y creación del mockup.

En el siguiente cuadro se presentan las principales actividades y la participación de los miembros del equipo en la etapa inicial del proyecto:

Actividades	**Lizeth García**	**Leonardo Guzmán**	**Alejandra Herrera**	**Carlos Silva**
Búsqueda de bases de datos para el desarrollo del proyecto	X	X	X	X
Identificación del problema y redacción del contexto	X	X		
Formulación de la pregunta de negocio y definición del alcance	X	X		
Descripción y documentación del conjunto de datos	X		X	
Análisis y exploración inicial de los datos		X		X
Elaboración del mockup		X		X
Consolidación y revisión de resultados		X		X