MALAGAHOUSE

Modelo predictivo de precios de viviendas en Málaga Capital

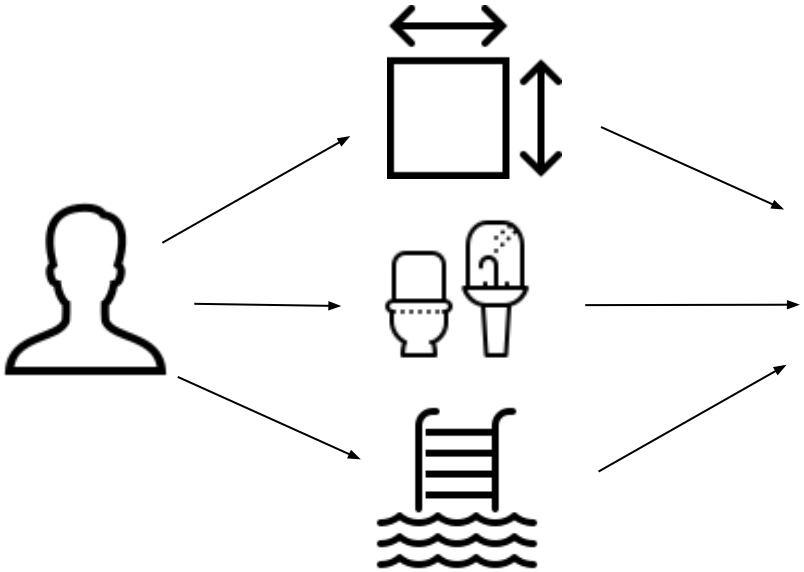
Realizado por Miguel Gámez y Sergio Toscano

Málaga Tech Park
Centro Público Integrado de FP

ÍNDICE

DESCRIPCIÓN

El proyecto tiene como finalidad dar un precio rápido a una vivienda para estudiar el mercado de las zonas de Málaga cómoda y rápidamente.

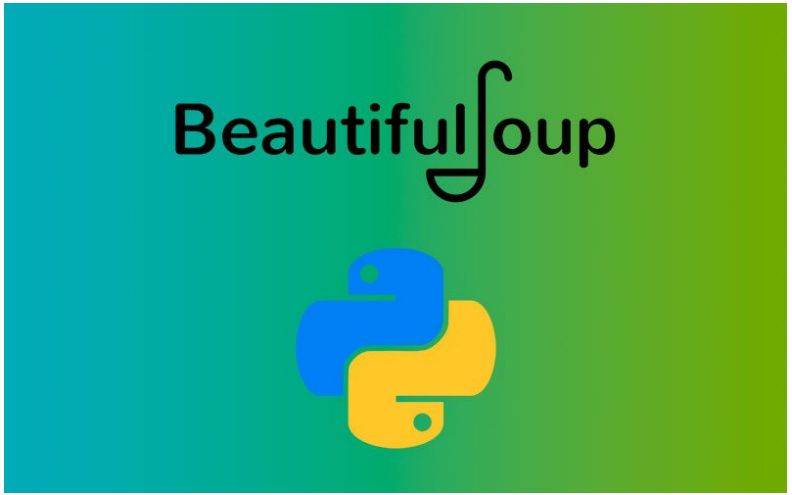


OBTENCIÓN DE DATOS

Web donde hemos obtenido los datos

pisos.com

Lenguaje y librería utilizados

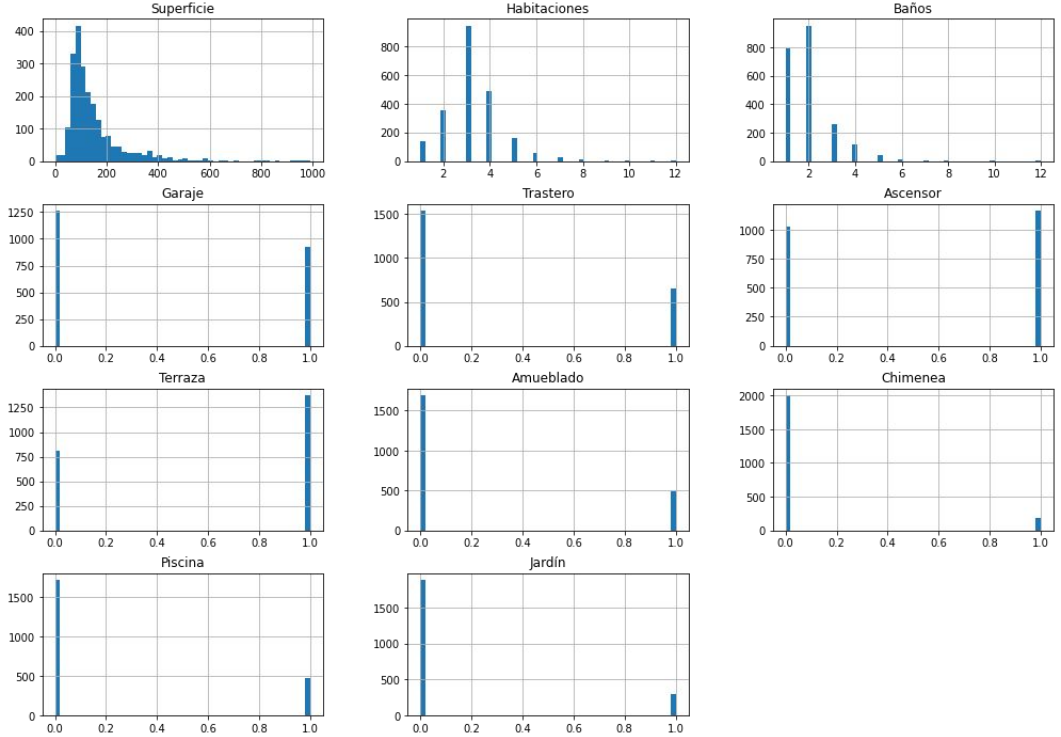


Código fuente:

EXPLORACIÓN Y VISUALIZACIÓN DE DATOS

```
[ ]  houses['Zona'].value_counts()

Centro                  501
Este                    331
Carretera de Cádiz      289
Bailén-Miraflores       223
Cruz de Humilladero     195
Puerto de la Torre      147
Ciudad Jardín           124
Churriana               117
Teatinos-Universidad    102
La Rosaleda-La Roca      95
Campanillas              66
Name: Zona, dtype: int64
```



Código fuente:

LIMPIEZA DE DATOS PRECIO “A CONSULTAR” Y VALORES NULOS

	Tipo	Zona	Precio	Superficie	Habi
67	Piso	Cruz de Humilladero	A consultar	87	
567	Piso	Puerto de la Torre	A consultar	103	
585	Piso	Puerto de la Torre	A consultar	90	
727	Piso	Puerto de la Torre	A consultar	95	
1269	Chalet	Puerto de la Torre	A consultar	532	
1608	Piso	Puerto de la Torre	A consultar	203	
1697	Piso	Puerto de la Torre	A consultar	71	
1839	Chalet	Carretera de Cádiz	A consultar	73	
1880	Piso	La Rosaleda-La Roca	A consultar	84	

	Tipo	Zona	Precio	Superficie
33	NaN	Centro	240000	175
66	NaN	La Rosaleda-La Roca	295000	40
70	NaN	Centro	499000	115
73	NaN	Centro	289000	65
129	NaN	Centro	750000	164
...	...	...	...	...
2012	NaN	Centro	519900	154
2054	NaN	Centro	190000	69
2092	NaN	Centro	295000	72
2099	NaN	Carretera de Cádiz	1618000	209
2150	NaN	La Rosaleda-La Roca	385000	61

Código fuente:

PREPARACIÓN DE DATOS CONVERSIÓN CATEGORÍA A NÚMERO

```
houses["Zona"].unique()
```

```
array(['Bailén-Miraflores', 'Centro', 'Este', 'Ciudad Jardín',
       'Carretera de Cádiz', 'Teatinos-Universidad', 'Churriana',
       'La Rosaleda-La Roca', 'Cruz de Humilladero', 'Campanillas',
       'Puerto de la Torre'], dtype=object)
```

```
zona_num = np.arange(len(houses["Zona"].unique()))
zona_num
```

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10])
```

```
houses["Zona"].replace(houses_zona, zona_num, inplace=True)
houses
```

	Tipo	Zona	Precio	Superficie	Habitaciones	Baños	Garaje
0	Piso	0	115000	69	2	1	0
1	Piso	1	295000	70	2	1	0
2	Ático	2	477000	170	6	2	0
3	Dúplex	1	1950000	190	3	3	1
4	Casa	3	126000	90	3	1	0

Código fuente:

PREPARACIÓN DE DATOS CONVERSIÓN CATEGORÍA A NÚMERO

```
houses["Tipo"].unique()
```

```
array(['Piso', 'Ático', 'Dúplex', 'Casa', 'Chalet', 'Finca rústica',
       'Estudio', 'Loft'], dtype=object)
```

```
tipo_num = np.arange(len(houses["Tipo"].unique()))
tipo_num
```

```
array([0, 1, 2, 3, 4, 5, 6, 7])
```

```
houses["Tipo"].replace(houses_tipo, tipo_num, inplace=True)
houses
```

	Tipo	Zona	Precio	Superficie	Habitaciones	Baños	Garaje
0	Piso	0	115000	69	2	1	0
1	Piso	1	295000	70	2	1	0
2	Ático	2	477000	170	6	2	0
3	Dúplex	1	1950000	190	3	3	1
4	Casa	3	126000	90	3	1	0

Código fuente:

PREPARACIÓN DE DATOS ELIMINACIÓN DE COLUMNAS

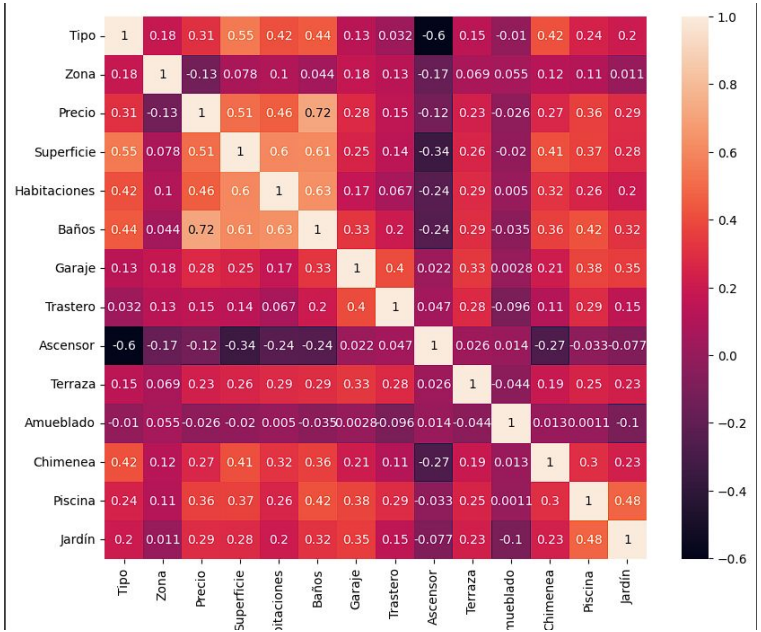
```
descripciones = houses
```

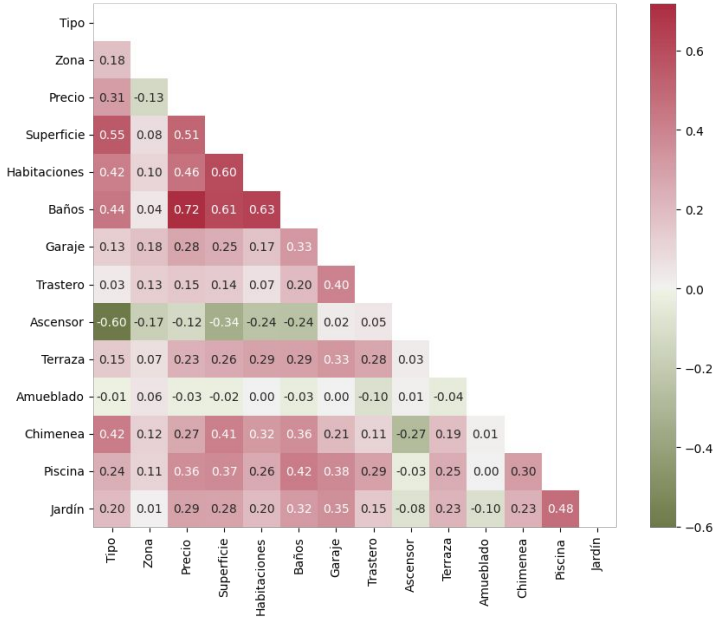
```
houses = houses.drop(columns='Descripción', axis=1)
houses
```

	Tipo	Zona	Precio	Superficie	Habitaciones	Baños	Garaje	Trastero	Ascensor	Terraza	Amueblado	Chimenea	Piscina	Jardín
0	0	0	115000	69	2	1	0	0	0	0	1	0	0	0
1	0	1	295000	70	2	1	0	1	1	0	0	0	0	0
2	1	2	477000	170	6	2	0	1	0	1	0	0	0	0
3	2	1	1950000	190	3	3	1	1	1	1	0	0	0	0
4	3	3	126000	90	3	1	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2185	0	0	160000	72	2	1	0	0	1	0	1	0	0	0
2186	0	1	182000	112	4	1	0	0	1	0	0	0	0	0
2187	3	6	320000	160	4	2	1	1	0	1	1	1	1	0
2188	0	5	340000	155	3	2	1	1	1	1	0	0	1	0
2189	3	7	650000	202	3	2	1	0	0	1	0	1	0	1

Código fuente:

PREPARACIÓN DE DATOS CORRELACIONES





Código fuente:

PREPARACIÓN DE DATOS

Como lo que queremos predecir es el Precio de las viviendas en Málaga, lo asignamos como dato de salida y (target). El resto como matriz de características X.

```
[137] y = houses['Precio']
      y

      0        115000
      1        295000
      2        477000
      3       1950000
      4        126000
               ...
      2185     160000
      2186     182000
      2187     320000
      2188     340000
      2189     650000
      Name: Precio, Length: 2094, dtype: int64
```

```
X = houses.drop('Precio', axis=1)
X
```

	Tipo	Zona	Superficie	Habitaciones	Baños	Garaje	Trastero	Ascensor	Terraza	Amueblado	Chimenea	Piscina	Jardín
0	0	0	69	2	1	0	0	0	0	1	0	0	0
1	0	1	70	2	1	0	1	1	0	0	0	0	0
2	1	2	170	6	2	0	1	0	1	0	0	0	0
3	2	1	190	3	3	1	1	1	1	0	0	0	0
4	3	3	90	3	1	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2185	0	0	72	2	1	0	0	1	0	1	0	0	0
2186	0	1	112	4	1	0	0	1	0	0	0	0	0
2187	3	6	160	4	2	1	1	0	1	1	1	1	0
2188	0	5	155	3	2	1	1	1	1	0	0	1	0
2189	3	7	202	3	2	1	0	0	1	0	1	0	1

2094 rows × 13 columns

Código fuente:

ENTRENAMIENTO DEL MODELO

Algoritmo	Error cuadrático medio	Coeficiente determinación
Linear Regression	304584.68	0.5608
Random Forest Regressor	220802.76	0.7225
Gradient Boosting Regressor	224323.68	0.7133
Bayesian Ridge	304429.91	0.5626
Cat Boost Regressor	227466.22	0.7817
XGBRegressor	225704.94	0.7436

Código fuente:

PROCESAMIENTO DE LENGUAJE NATURAL

Hemos decidido hacer una aplicación de **detección de entidades nombradas**. El objetivo es buscar todas las palabras o frases detectadas como “LOC” es decir, **localizaciones**. Con esto queremos ver los sitios más mencionados dentro de las **descripciones** extraídas.

Librería: Spacy

Código fuente:

DEMO

Streamlit

Página web

POSIBLES MEJORAS

- Entrenar el modelo con un volumen más amplio de datos.
- Hacer el modelo para que sirva en distintas zonas de España, y no solo de Málaga Capital.
- Incluir más campos relevantes como la planta de la vivienda (si tiene).
- Mejoras visuales para la aplicación web.

CONCLUSIONES

- La vivienda es cara.

- Algoritmos como CatBoostRegressor y XGBRegressor funcionan mejor en este tipo de problemas.

- Hemos podido ver las zonas más ofertadas y de más interés general como la malagueta, el limonar, etc.

BIBLIOGRAFÍA

- Temario de IA, MIA y Big Data.
- https://docs.streamlit.io/
- https://openai.com/blog/chatgpt
- https://www.pisos.com/
- https://youtu.be/D57kiTBFu3I

FIN

PROYECTO REALIZADO POR MIGUEL GÁMEZ RUIZ Y SERGIO TOSCANO DÍAZ