# Predicción del Precio de Viviendas mediante Regresión Lineal Múltiple

Autores: Apellidos Nombres1, Apellidos Nombres2, Apellidos Nombres3, Apellidos Nombres4

1correo institucional, 2correo institucional, 3correo institucional, 4correo institucional

Carreras: 1Ingeniería\_\_\_, 2 Ingeniería\_\_\_, 3 Ingeniería\_\_\_, 4Ingeniería\_\_\_

Asesor: Dra. Jessica Briseño

## 1. Resumen

El presente trabajo tiene como finalidad aplicar un modelo de regresión lineal múltiple para predecir el precio de una casa (en dólares) a partir de tres variables: la superficie total de la casa en pies cuadrados, el número de habitaciones y la antigüedad de la vivienda. Se emplearon herramientas estadísticas y el lenguaje de programación Python, para realizar los análisis y poner en práctica los conocimientos adquiridos en la asignatura de Estadística Descriptiva. Los resultados obtenidos muestran una relación significativa entre las variables independientes y el precio de venta de las viviendas.

## 2. Introducción

El análisis del mercado inmobiliario es una herramienta tanto para compradores como para vendedores. En este contexto, la estadística permite construir modelos predictivos que ayudan a estimar el valor de una propiedad con base en sus características físicas (REFERENCIA). Este estudio se enfoca en la aplicación del método conocido como regresión lineal múltiple para predecir el precio de venta de casas en Estados Unidos, utilizando como variables predictoras el tamaño del primer piso, el número de habitaciones y la antigüedad de la vivienda.

## 3. Objetivos del estudio

El objetivo de este trabajo es generar un modelo de regresión lineal múltiple para predecir el precio de las viviendas (en dólares) en función de tres variables: superficie (m²), número de habitaciones y antigüedad (años), aplicando los conocimientos de Estadística Descriptiva como base para futuros modelos de estimación con inteligencia artificial.

## 4. Hipótesis

Existe una relación significativa entre el precio de venta de una casa y las variables superficie de la vivienda, número de habitaciones y antigüedad.

## 5. Metodología

La regresión lineal es uno de los algoritmos más conocidos y comprendidos tanto en estadística como en aprendizaje automático supervisado, lo que permite considerarlo una forma fundamental de inteligencia artificial (Brownlee, 2020). En este estudio, se utilizó el método de regresión lineal múltiple, el cual permite estimar el valor de una variable dependiente () en función de varias variables independientes (). La ecuación general del modelo de regresión múltiple es (REFERENCIA):

Donde:  
: Precio de venta de la casa (Variable dependiente)

Variables independientes  
: Superficie de la casa en pies cuadrados  
: Número de habitaciones  
: Antigüedad de la casa en años  
: Intercepto  
: Coeficientes de regresión  
: Término de error

Los coeficientes de regresión se calculan mediante la ecuación matricial de mínimos cuadrados que es:

Donde:

es la matriz de las variables independientes. La primera columna de esta matriz contiene unos.

es el vector de salida

es el vector de coeficientes de regresión

El análisis fue realizado utilizando el lenguaje de programación Python (Rossum, 1991) en el entorno colaborativo Google Colab, haciendo uso de bibliotecas como pandas, numpy, matplotlib, seaborn y statsmodels para la manipulación de datos, visualización y cálculo del modelo de regresión.

## 6. Desarrollo

La base de datos (dataset) utilizada fue descargada de la plataforma Kaggle y lleva por nombre House Prices - Advanced Regression Techniques (Kaggle, n.d.). Para este análisis, se seleccionaron únicamente tres columnas representativas. La Tabla 1 muestra el nombre de la variable utilizada en el análisis, el nombre correspondiente en el dataset y una breve descripción de cada una.

|  |  |  |
| --- | --- | --- |
| Variable | Columna | Indicador español |
| Y | SalePrice | Precio de venta de la casa en dólares |
| X1 | TotSF | Superficie Total de la casa |
| X2 | BedroomAbvGr | Número de habitaciones |
| X3 | AntiqueYear | Edad de la casa en Años |

**Tabla 1. Variables empleadas en el estudio**

El análisis consistió en los siguientes pasos:

* Carga del archivo CSV desde Kaggle.
* Selección de variables relevantes.
* Cálculo de estadísticas descriptivas básicas.
* Generación de histogramas
* Generación del modelo de regresión lineal múltiple para obtener la ecuación predictiva.

Todo el análisis se realizó en Python, facilitando el procesamiento de datos y la visualización de resultados.

## 7. Resultados

La tabla 2 muestra los resultados de los principales estadísticos como son el número de datos (count), la media (mean), la desviación estándar (std) el valor mínimo (min), los cuartiles 1, 2 y 3 (25%, 50% y 75%), así como el valor máximo para cada variable.

Tabla

El contenido generado por IA puede ser incorrecto.

**Tabla 2. Principales estadísticos de las variables de estudio**

El análisis de la distribución de las variables incluidas en el modelo se llevó a cabo mediante la generación de histogramas, utilizando la biblioteca matplotlib de Python. En la Figura 2 muestra los histogramas correspondientes a las variables consideradas en el modelo de regresión lineal múltiple: precio de venta (SalePrice), superficie total en pies² (TotSF), el número de habitaciones (BedroomAbvGr) y antigüedad de la casa (AntiqueYear).

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

**Figura 1. Histográmas de las variables de estudio**

El análisis gráfico de estas distribuciones permite identificar patrones y posibles sesgos en los datos, los cuales pueden influir en la calidad del modelo predictivo.

* La variable *SalePrice*, que representa el precio de venta de las casas, muestra una distribución sesgada a la derecha, concentrándose entre 150,000 y 200,000 dólares, con una disminución progresiva en la frecuencia a medida que aumenta el precio. Esto sugiere la presencia de valores atípicos que podrían afectar el modelo si no se mitigan.
* La variable *TotSF* (superficie del primer piso) tiene una distribución asimétrica positiva, con la mayoría de las viviendas entre 1,000 y 2,000 pies², y algunos casos con más de 3,000 pies². Esta dispersión refleja la variedad en el tamaño de las casas, lo que podría generar sensibilidad a los valores extremos.
* La variable *BedroomAbvGr* (número de habitaciones) presenta una distribución discreta, concentrada en 3 habitaciones, con pocos casos de más de 5. Esto refleja una tendencia en el diseño estándar de viviendas familiares.
* La variable *AntiqueYear* (antigüedad de la casa) muestra una distribución asimétrica negativa, con muchas casas entre 20 y 40 años, pero también algunas más antiguas, indicando una mezcla de propiedades nuevas y más históricas en el conjunto de datos.

El modelo de regresión lineal múltiple se implementó utilizando el lenguaje de programación Python en el entorno colaborativo de Google Colab. El código desarrollado para su ejecución se incluye en el anexo de este trabajo. A continuación, se presentan los resultados obtenidos del modelo de regresión estimado:

Donde:  
: Precio de venta de la casa  
: Superficie de la casa en pies cuadrados  
: Número de habitaciones  
: Antigüedad de la casa en años

El modelo indica que:

* El valor de $ es el precio base de las casas.
* Cada pie² adicional en la superficie de la casa, el precio incrementa en promedio $8.10 dólares
* Cada habitación adicional aumenta el precio en $13,197.82 dólares.
* Cada año adicional de antigüedad incrementa el precio en $21.37 dólares.

El coeficiente de determinación tiene un valor de R²=0.655, lo que indica que aproximadamente el 65.52% de la variabilidad en el precio de venta de las casas puede explicarse por las variables seleccionadas.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

**Figura 2. Relación entre el precio de venta y las variables explicativas del modelo**

La figura 2 muestra tres diagramas de dispersión que ilustran la relación individual entre el precio de venta de las casas (variable dependiente) y cada una de las variables explicativas consideradas en el modelo: superficie de la casa, número de habitaciones y edad de la vivienda. En cada gráfico se ha trazado la línea de regresión correspondiente, lo que permite visualizar la tendencia general de los datos.

En la primera gráfica, correspondiente a la relación entre el precio de venta y el tamaño de la casa (TotSF), se observa una tendencia positiva, lo cual indica que a medida que incrementa el tamaño de la vivienda, el precio tiende también a aumentar. La segunda gráfica representa la relación entre el precio de venta y el número de habitaciones (BedroomAbvGr). En este caso también se evidencia una tendencia ascendente, lo que sugiere que un mayor número de habitaciones está asociado con un incremento en el precio de la vivienda. Finalmente, en la tercera gráfica se muestra la relación entre el precio de venta y la antigüedad de la casa (AntiqueYear). A diferencia de las variables anteriores, en este caso la tendencia es mucho más débil. La línea de regresión muestra una pendiente casi nula, lo que indica que la edad de la vivienda no tiene una influencia tan marcada en el precio como las otras variables.

## 8. Conclusiones

A partir del análisis realizado, se puede concluir que existe una relación significativa entre el precio de venta de las casas y las variables superficie de la vivienda (pies²), número de habitaciones y antigüedad. El modelo de regresión lineal múltiple desarrollado en este trabajo resultó ser útil para estimar el precio de venta de las viviendas en función de su tamaño, número de habitaciones y antigüedad. El coeficiente de determinación obtenido sugiere una buena capacidad predictiva del modelo; sin embargo, aún existe margen para mejorar la predicción mediante la inclusión de otras variables, como ubicación, acabados, entre otras.

La superficie y el número de habitaciones presentan una correlación positiva con el precio, lo que indica que viviendas más grandes y con más habitaciones tienden a tener precios más elevados. En contraste, la antigüedad de la casa muestra una relación negativa con el precio, lo que sugiere que las viviendas más nuevas tienen un valor más alto en promedio. Estos resultados respaldan la hipótesis planteada, aunque es importante considerar otros factores adicionales que también podrían influir en el precio de venta.

Finalmente, se logró aplicar de manera satisfactoria el enfoque de regresión en un entorno computacional utilizando herramientas estadísticas, lo que contribuyó a consolidar los conocimientos adquiridos en la asignatura de Estadística Descriptiva.

## Referencias

Brownlee, J. (20 de Abril de 2020). *Linear Regression for Machine Learning. Machine Learning Mastery.* . Obtenido de https://machinelearningmastery.com/linear-regression-for-machine-learning/

Kaggle. (n.d.). *House Prices - Advanced Regression Techniques [Dataset].* Obtenido de https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

Rossum, G. v. (1991). *python*. Obtenido de https://www.python.org/

## Integridad Académica

Este trabajo fue realizado de manera individual y respetando los principios de integridad académica. Se utilizó información pública y se citó adecuadamente la fuente de la base de datos empleada. Todas las herramientas utilizadas fueron de acceso libre y con fines educativos.

## Anexo. Código

Texto, Carta

El contenido generado por IA puede ser incorrecto.