ANÁLISIS DE SERIES TEMPORALES Y OPTIMIZACIÓN DE MODELOS MEDIANTE ALGORITMOS DE OPTIMIZACIÓN

INTRODUCCIÓN A LA IA

ENTREGA FINAL

Yeniffer Andrea Córdoba
Pablo Gómez Mutis
Luz Adriana Yepes

SERIES DE TIEMPO

Las series de tiempo son un conjunto de datos secuenciales que se recopilan o registran en **intervalos de tiempo** específicos, y se utilizan para analizar patrones, tendencias y pronósticos a lo largo del tiempo.

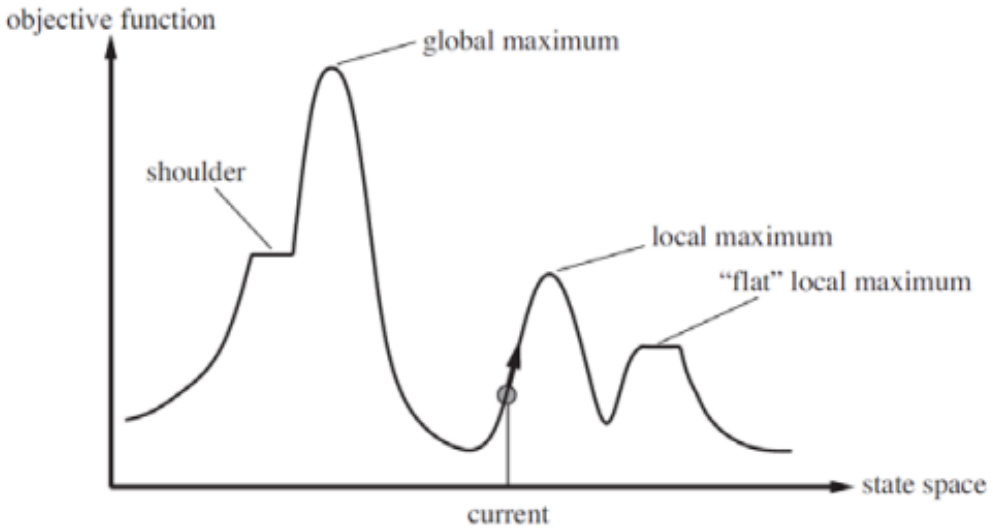
Modelos estadísticos tradicionales → ARIMA
Modelos estadísticos tradicionales → VAR

Modelos *Machine Learning* → Random Forest

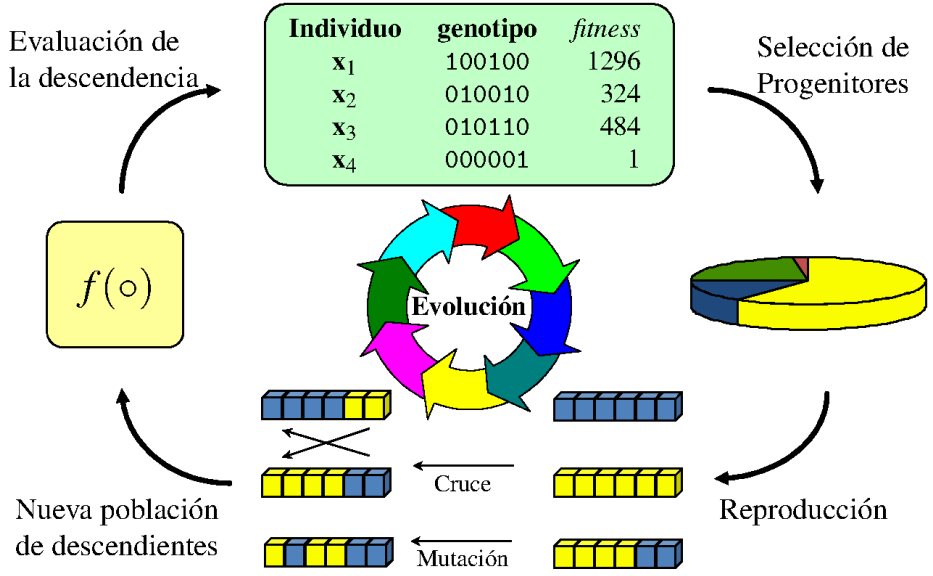
ALGORITMOS DE OPTIMIZACIÓN

Los algoritmos de optimización son métodos matemáticos o computacionales utilizados para encontrar la mejor solución posible a un problema, ajustando las variables de entrada para maximizar o minimizar una función objetivo bajo ciertas restricciones.

Hill Climbing



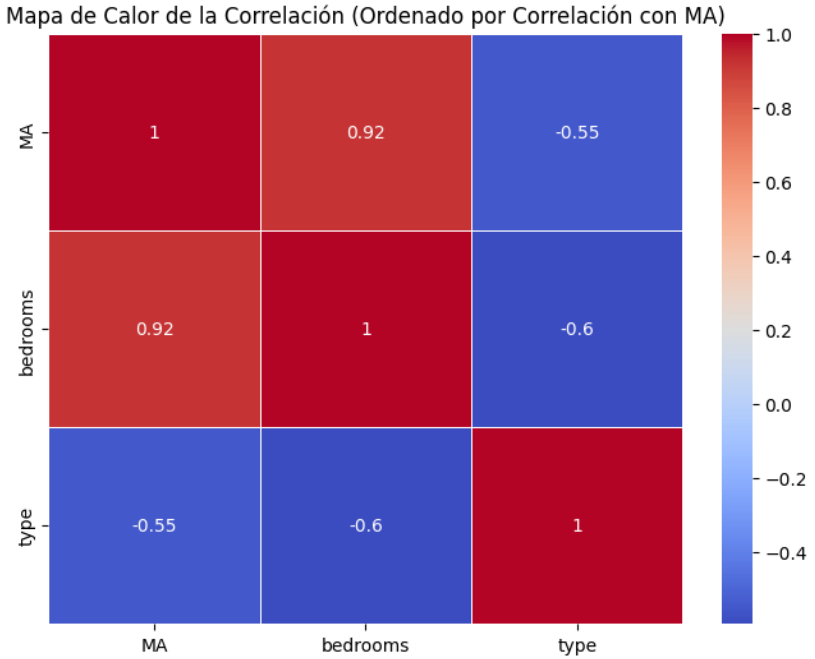
Algoritmo Genético

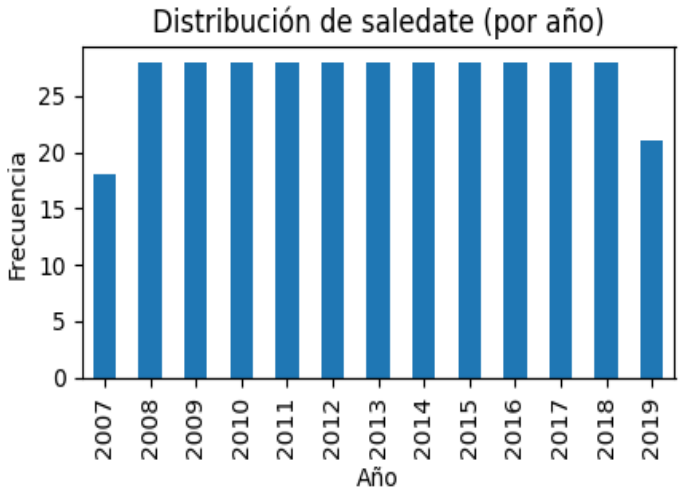


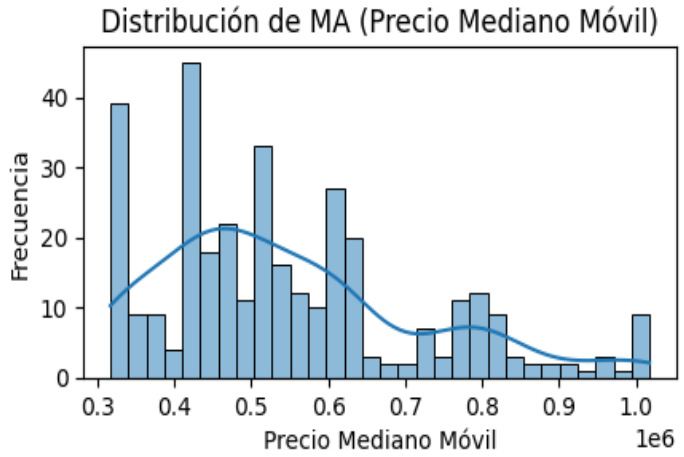

DATASET
PREPROCESAMIENTO
ARIMA
VAR
RANDOM FOREST
HILL CLIMBING
HILL CLIMBING
HILL CLIMBING
ALG. GENÉTICO
ALG. GENÉTICO
ALG. GENÉTICO
COMPARACIÓN DE RESULTADOS

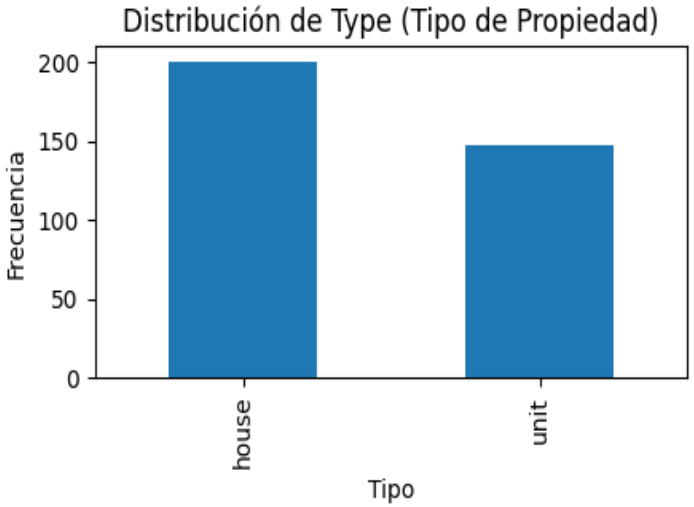
DATASET

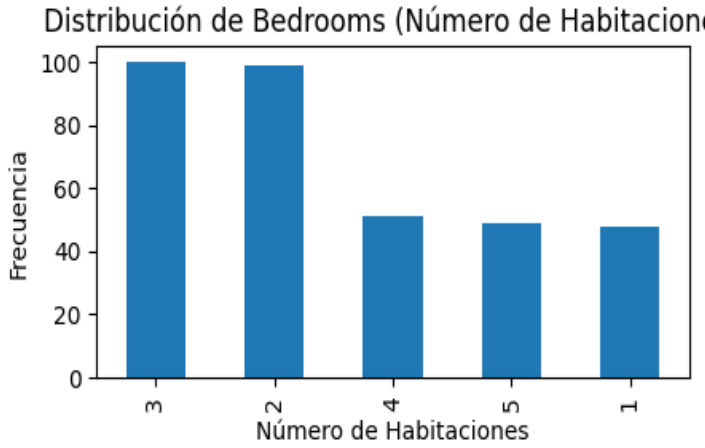
	saledate	MA	type	bedrooms
0	30/09/2007	441854	house	2
1	31/12/2007	441854	house	2
2	31/03/2008	441854	house	2











ARIMA (Autoregressive Integrated Moving Average)

ARIMA es un modelo estadístico **univariado** utilizado para predecir series de tiempo, combinando tres componentes:

Autorregresivo (AR)

Utiliza valores pasados de la serie para predecir el valor actual, basándose en la relación lineal entre observaciones anteriores y la observación actual.

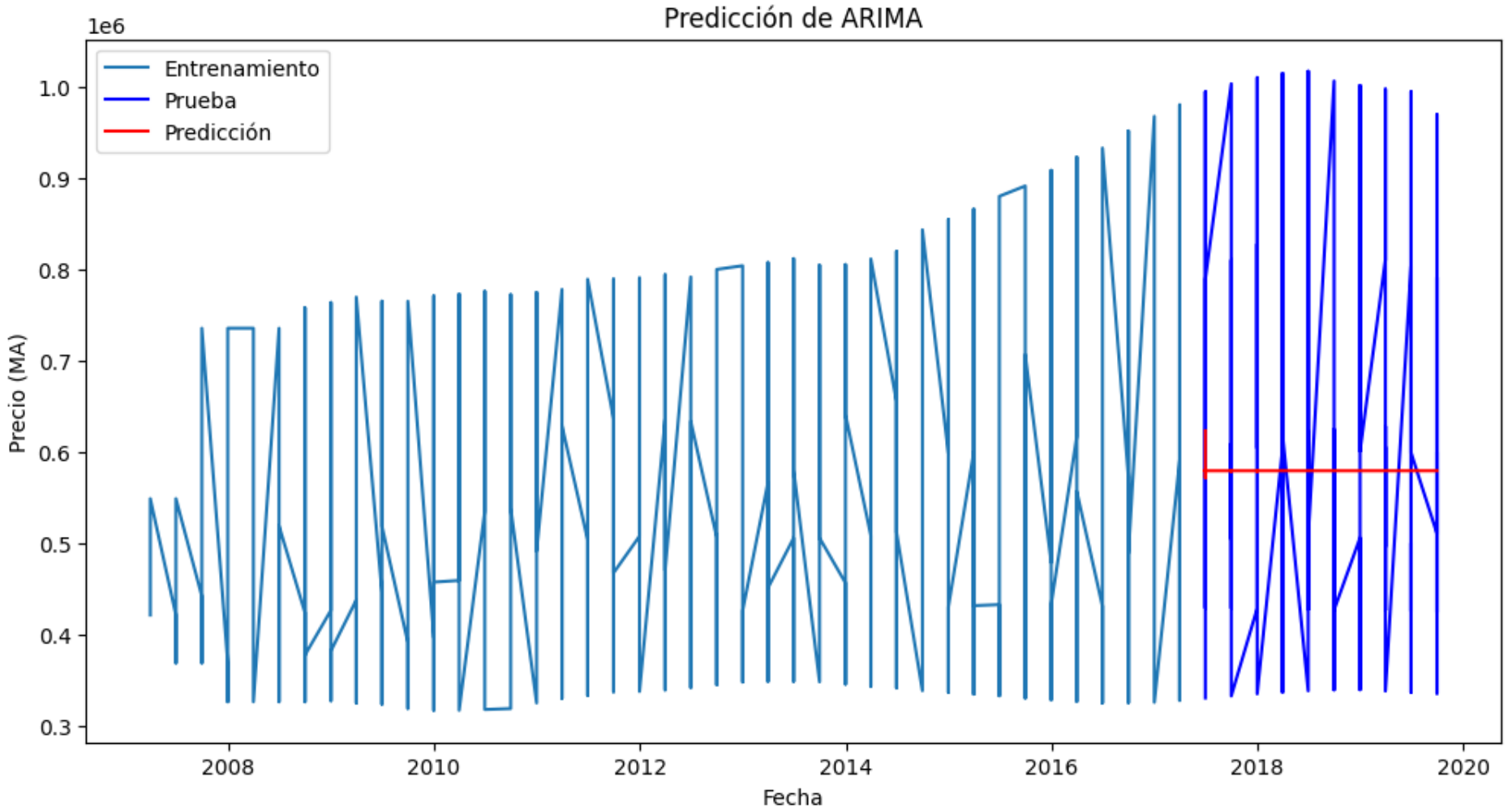
Integrado (I)

Aplica diferenciación para hacer que la serie sea estacionaria, eliminando tendencias o estacionalidades que puedan distorsionar las predicciones.

Media Móvil (MA)

Modela la relación entre el valor actual y los errores de predicción previos, ajustando las predicciones con base en errores pasados.

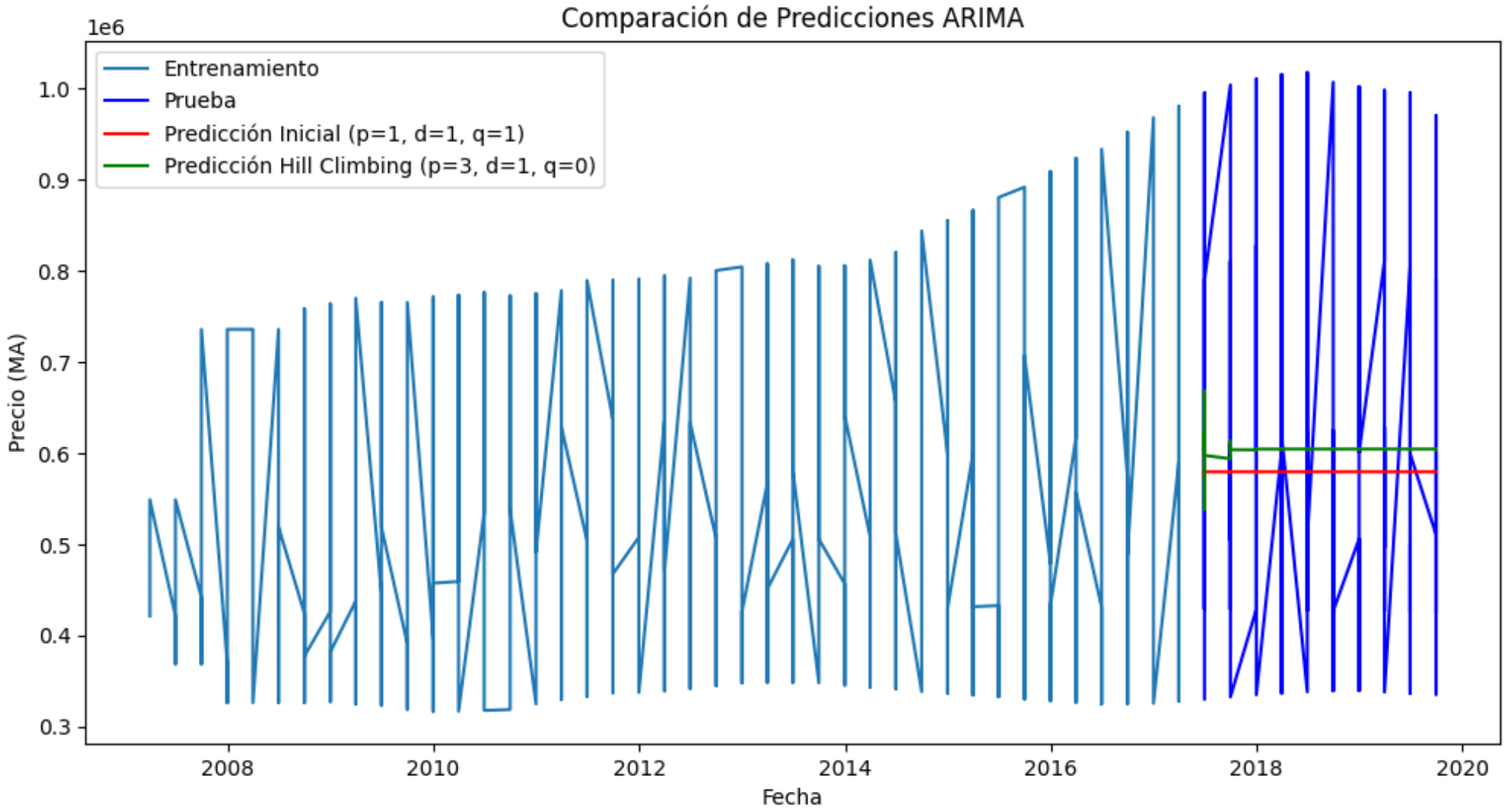
MODELO ORIGINAL



MSE: 45260747979.29171
RMSE: 212,745.7355137623

Parámetros iniciales: p=1, d=1, q=1

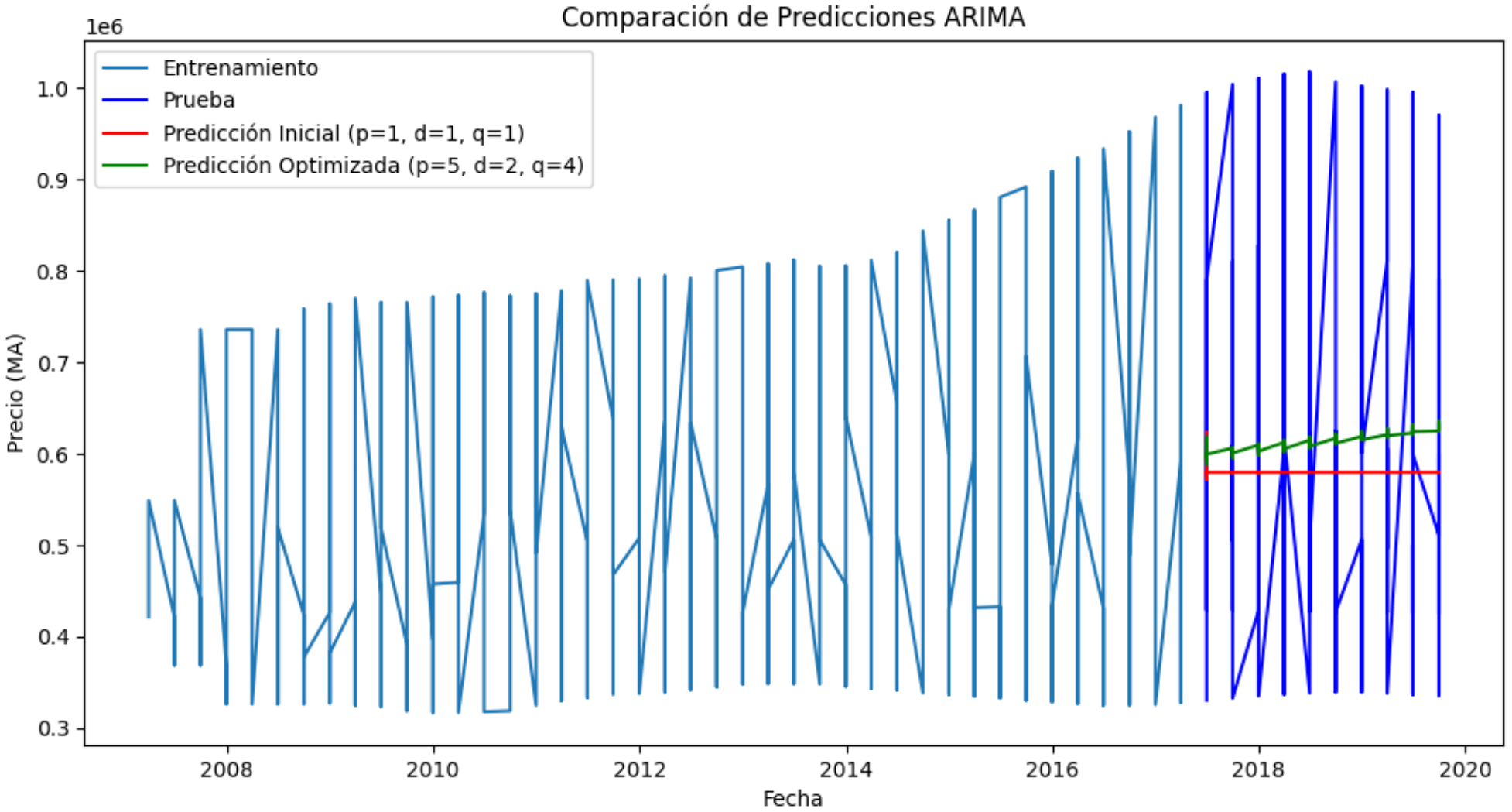
HILL CLIMBING



MSE: 45089717328.55228
RMSE: 187’669,5

Parámetros iniciales: p=3, d=1, q=0

ALGORITMO GENÉTICO



MSE: 45134299511.83591
RMSE: 212,448.34551447065

Parámetros iniciales: p=5, d=2, q=4

VAR (Vector Autoregressive)

VAR es un modelo estadístico utilizado en series de tiempo **multivariadas**, que puede capturar las relaciones entre múltiples variables dentro de un mismo modelo, considerando las siguientes características:

Relaciones Cruzadas entre variables

Cada variable se modela en función de sus propios rezagos y los rezagos de las otras variables.

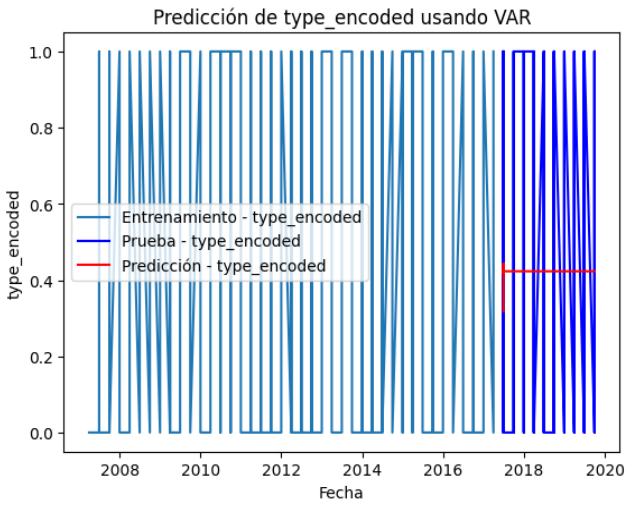
Estacionariedad

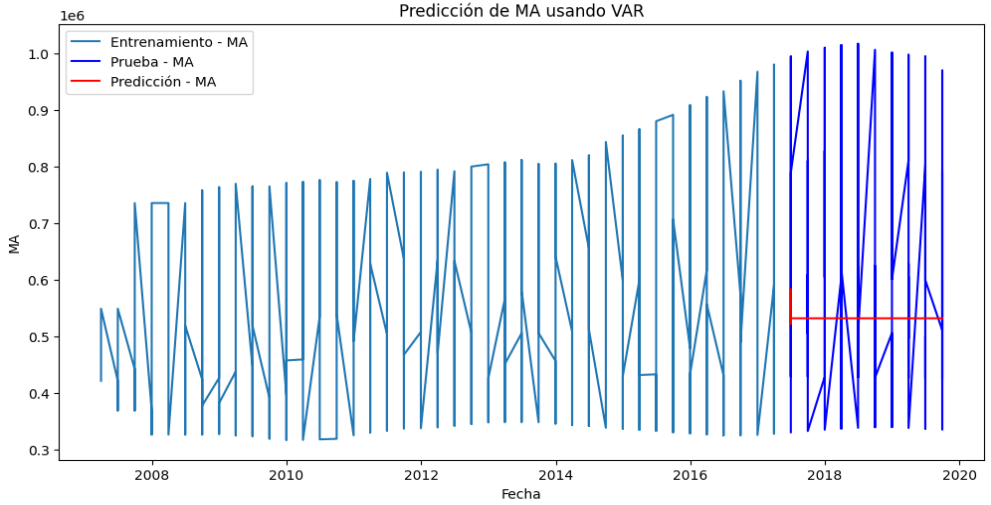
Análogo al modelo ARIMA, las características estadísticas deben ser constantes en el tiempo antes de aplicar el modelo.

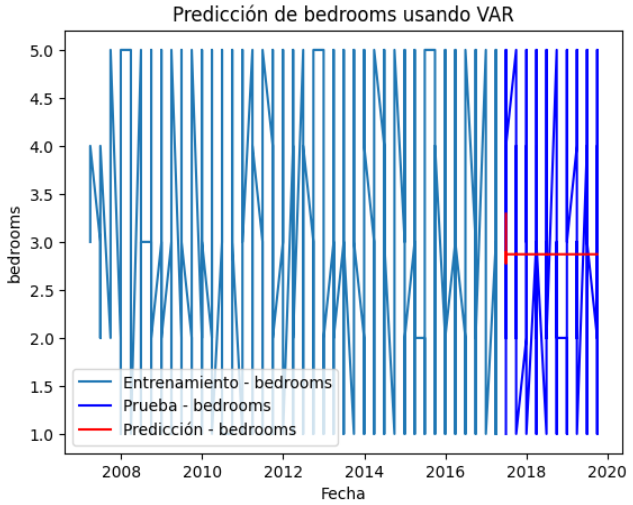
Orden del Modelo (Lags)

Este parámetro determina cuántos períodos pasados de cada variable se deben considerar para predecir el valor actual.

MODELO ORIGINAL







type_encoded

MSE: 0,24698
RMSE: 0,49697

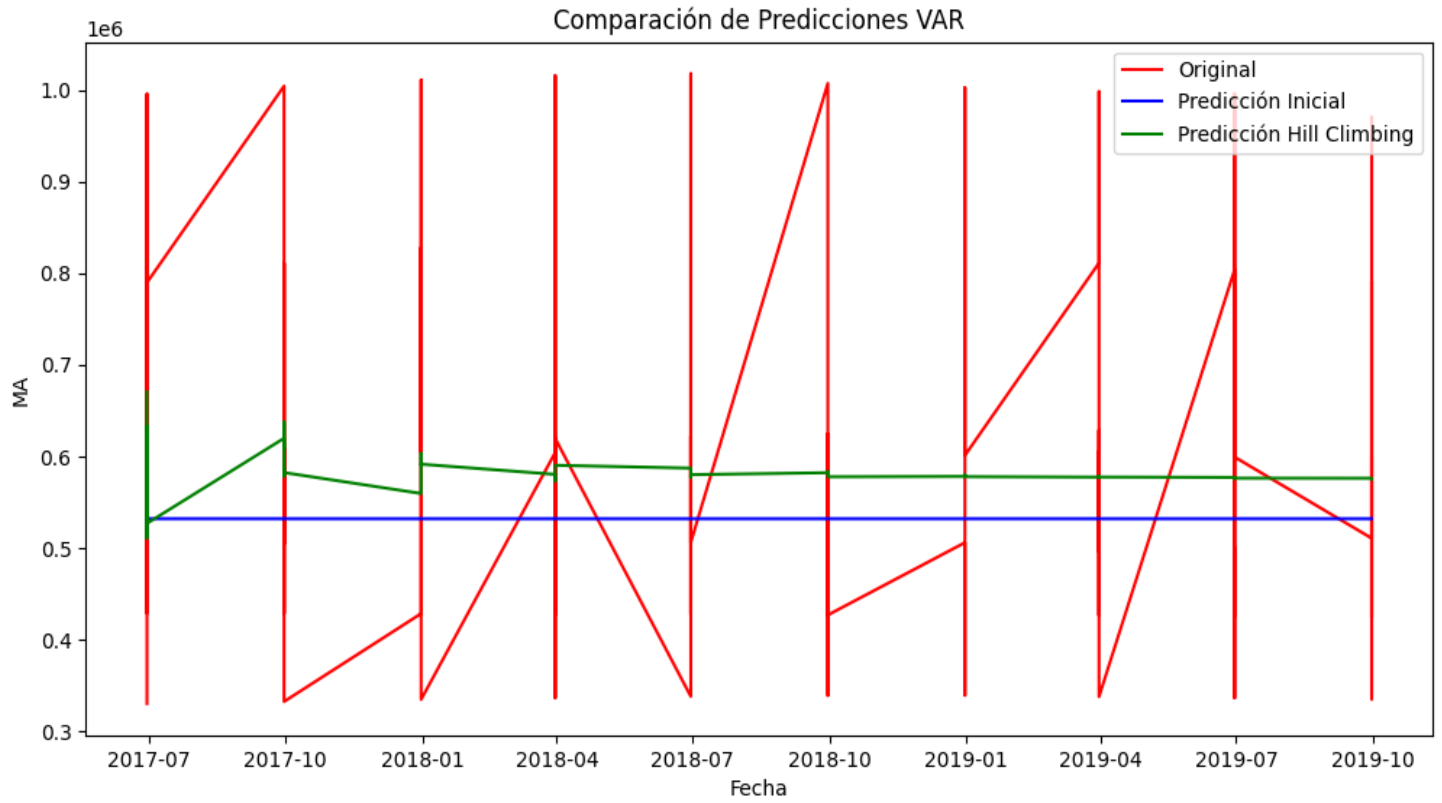
Precio medio (MA)

MSE: 52.420.292.804,45083
RMSE: 228.954,78331

Bedrooms

MSE: 1,57253
RMSE: 1,25400

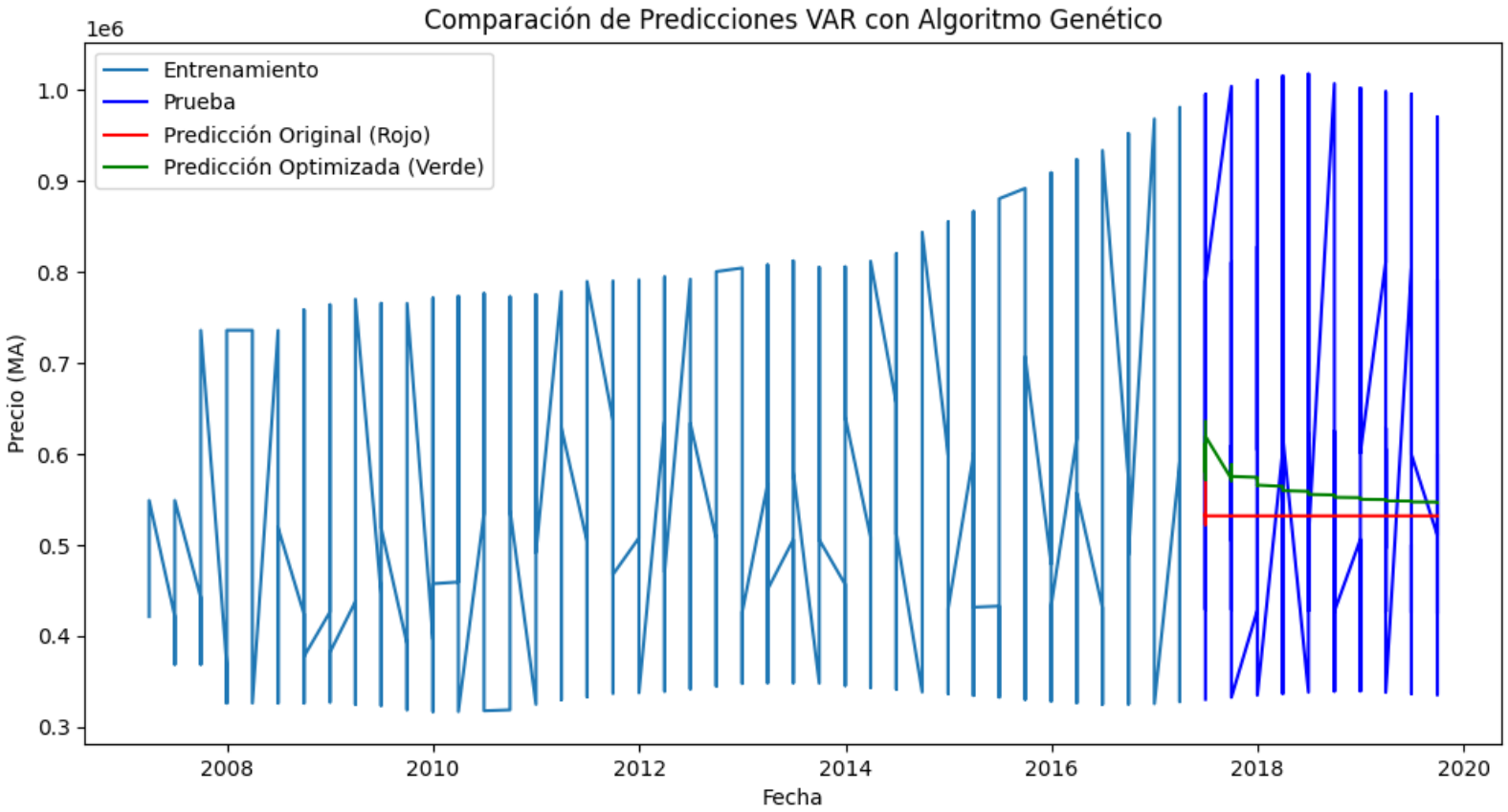
HILL CLIMBING



Precio medio (MA)

MSE: 23.353.468.550,93337
RMSE: 152.818,41692

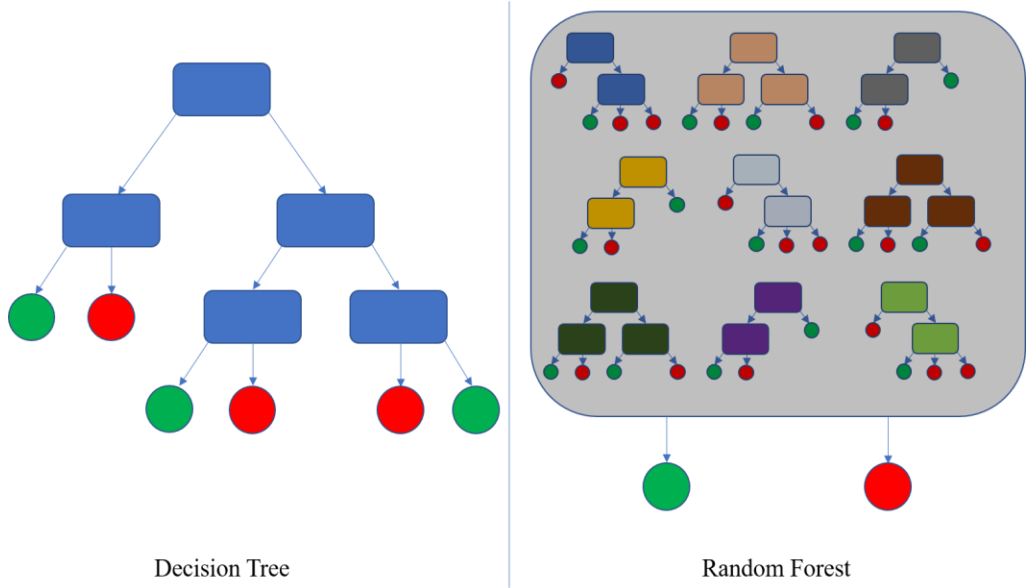
ALGORITMO GENÉTICO



MSE: 48.245.819.042,45697
RMSE: 219.649,30922

RANDOM FOREST

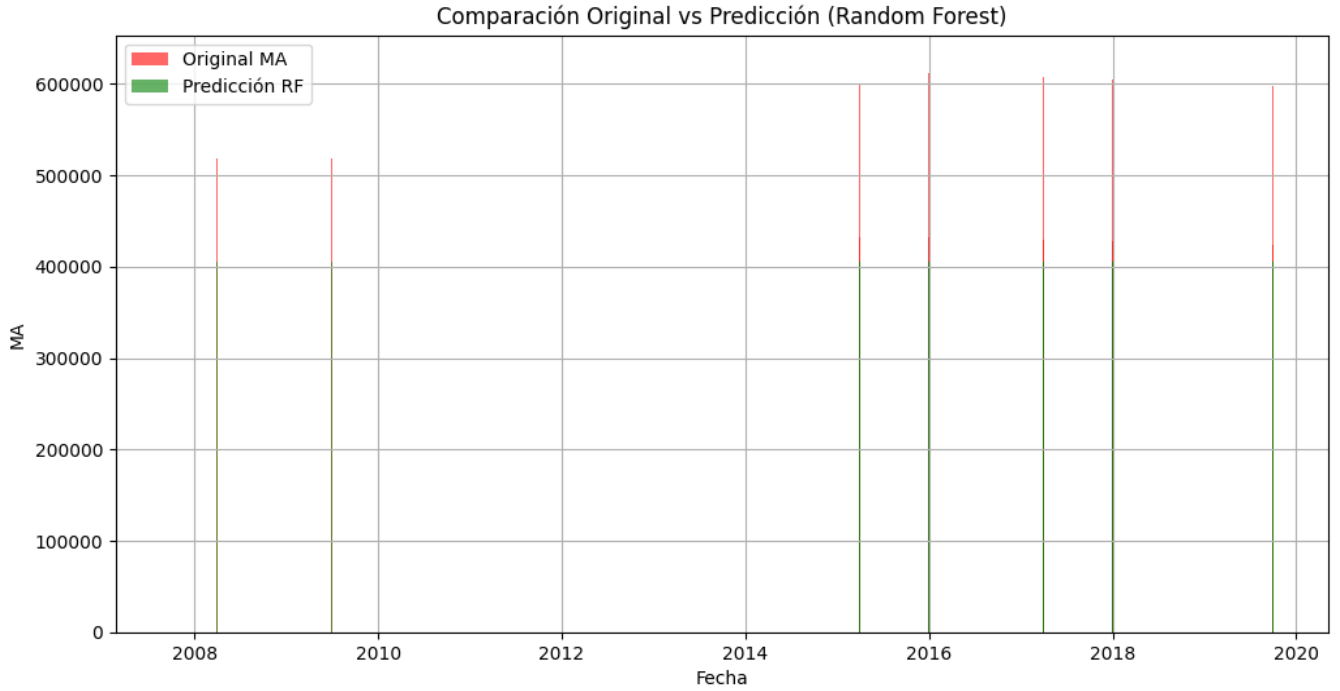
Este algoritmo de Machine Learning se basa en la construcción de múltiples árboles de decisión para realizar predicciones, mejorando la precisión de los modelos individuales al combinar los resultados de varios árboles, reduciendo el sobreajuste.



Para entrenar cada árbol, este algoritmo utiliza un subconjunto aleatorio de los datos , empleando una técnica llamada ***bagging*** (bootstrap aggregating).

En nuestro caso, como la predicción de precios de vivienda es un problema de regresión, el modelo toma el promedio de las predicciones de todos los árboles, lo que aumenta la precisión general del modelo.

MODELO ORIGINAL



MSE: 19.880.036.499,78398 **RMSE**: 140.996,58329 **MAE**: 121.783,76866

HILL CLIMBING

El algoritmo Hill Climbing optimizó el modelo de Random Forest con los siguientes parámetros:

- n_estimators = 65
- max_depth = None

El resultado obtenido es un **MSE** de 19,777,496,933.43, lo que produce un **RMSE** de 140,632.5.

Esto significa que hubo una mejora del error de 0.26%.

ALGORITMO GENÉTICO

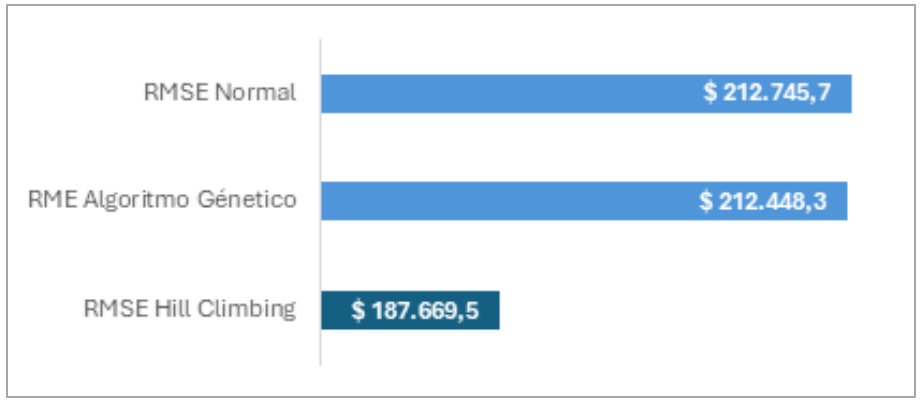
El algoritmo Genético optimizó el modelo de Random Forest en la Generación 50 con el siguiente resultado:

- **Mejor MSE** = 19,690,182,510.28
- **RMSE** = 140,235.35

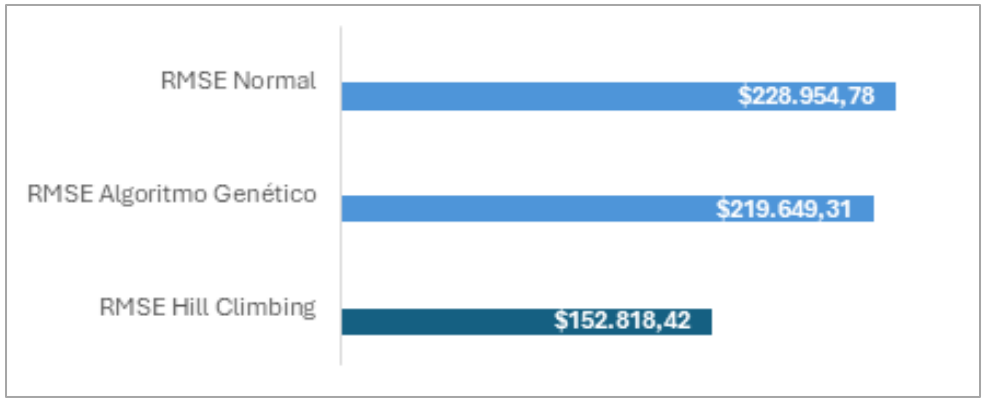
El algoritmo Genético logró una mejora del 0.54% en la reducción del **MSE** y **RMSE**.

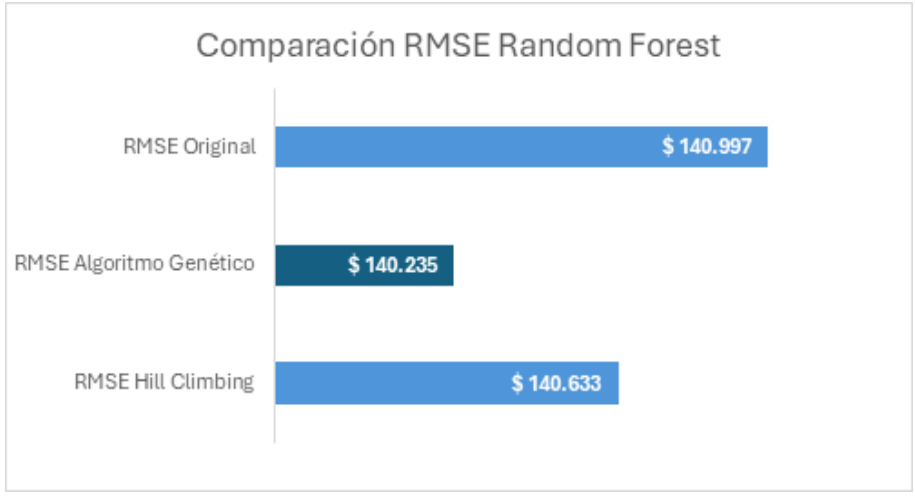
CONCLUSIONES

Comparación ARIMA

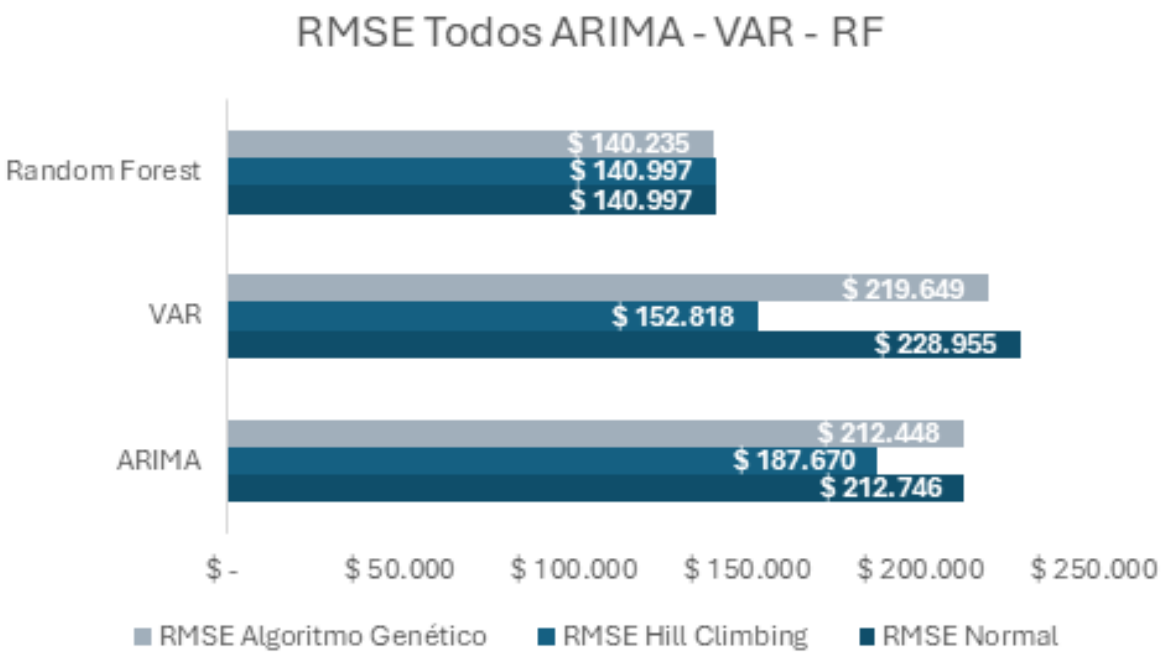


Comparación VAR





CONCLUSIONES



Cambio (%)	ARIMA	VAR	Random Forest
RMSE Algoritmo Genético	0,14%	4%	0,54%
RMSE Hill Climbing	11,8%	33%	0,26%