UNIVERSIDAD EAFIT

PREDICCIÓN DE PRECIOS DEL SECTOR INMOBILIARIO EN ZONAS COSTERAS DEL ATLANTICO EN EE. UU., MEDIANTE EL USO DE TÉCNICAS DE MACHINE LEARNING.

Prediction of real estate prices in Atlantic coastal areas in the USA, using machine learning techniques.

SARA GALLEGO MUÑOZ

Proyecto de Grado

Asesor, docente
Paula María Almonacid Hurtado

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA
MEDELLÍN
2024

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ECUACIONES

RESUMEN

El sector de bienes inmuebles es fundamental para las economías, representando un significativo porcentaje del PIB en la economía de países desarrollados y emergentes. Este mercado involucra actores clave como inversionistas, compradores, entidades financieras e instituciones gubernamentales, quienes requieren predicciones precisas sobre los precios medianos de los inmuebles para ajustar estrategias, políticas y toma de decisiones informadas.

En los últimos años se han propuesto metodologías para estimar el precio de bienes inmuebles en ciudades urbanas centrales, usando variables asociadas a la descripción del inmueble y geográficas de la ubicación. En este trabajo se propone una metodología que permita predecir el precio mediano de bienes inmuebles en zonas costeras del Atlántico en EE. UU., haciendo uso de variables relacionadas tanto con el inmueble como con su entorno y la aplicación de técnicas de Machine Learning y Deep Learning. Se espera obtener la predicción del precio de bienes inmuebles para zonas costeras del Atlántico en EE. UU., de las cuales no se suele obtener esta valoración de manera pública y confiable.

Palabras claves: *mercado inmobiliario, machine learning, zona costera* atlántica

ABSTRACT

The real estate sector is fundamental to economies, representing a significant percentage of GDP in the economies of developed and emerging countries. This market involves key players such as investors, buyers, financial entities and government institutions, who require accurate forecasts on real estate median prices to adjust strategies, policies and make informed decisions.

In recent years, methodologies have been proposed to estimate the price of real estate in central urban cities, using variables associated with the property description and geographic location. In this paper we propose a methodology to predict the median price of real estate in Atlantic coastal areas in the U.S., using variables related to both the property and its environment and the application of Machine Learning and Deep Learning techniques. It is expected to obtain the prediction of the price of real estate for Atlantic coastal areas in the U.S., for which this valuation is not usually obtained in a public and reliable way.

Keywords: *real estate market, machine learning, Atlantic coastal area.*

1. INTRODUCCIÓN

El sector de los bienes inmuebles compone entre el 3% y el 5% del PIB en países como EE.UUU [1], y el 25% en China [2]. Es decir, las transacciones asociadas a la compra y venta de inmuebles son importantes para las economías, tanto para las desarrolladas como las emergentes.

En el mercado inmobiliario, interactúan actores clave como inversionistas, compradores; quienes se interesan en la predicción de los precios medianos de bienes inmuebles, con el objetivo de ajustar sus estrategias de compra y venta a través de la toma de decisiones informadas sobre los precios de las viviendas. Otro actores interesados son las entidades financieras e instituciones gubernamentales, quienes se benefician de esta predicción al proveerles información para el desarrollo de políticas de vivienda, programas de desarrollo económico y planificación urbana, y a entender cómo factores como el empleo y los ingresos influyen en los precios de las viviendas.

La información necesaria para realizar análisis predictivos sobre los precios de los inmuebles en el mercado inmobiliario no es de usual recolección, y dada su recolección suele ser poco precisa. Adicionalmente, se han propuesto metodologías para estimar el precio de bienes inmuebles, usando información descriptiva del entorno y estructura de los inmuebles; pero pocas investigaciones sugieren metodologías para estimar dichos valores en ciudades con zonas costeras y otras variables relacionadas con este tipo de ciudades.

En este trabajo se plantea una metodología que busca predecir el precio mediano de bienes inmuebles ubicados en la zona costera Atlántica de EE. UU., utilizando variables relacionadas con su entorno, en un contexto socio económico y climático, asociandolas a las características del territorio en el que se encuentren, climático. Esto se lleva a cabo a partir del uso de algoritmos de Machine Learning (ML) especializados en regresión que permiten realizar la predicción de precios; adicional a usar técnicas que permiten determinar las variables más importantes para la predicción, permitiendo un mayor entendimiento del comportamiento del mercado. ,

En la primera sección de este documento,se incluye una explicación más amplia del problema a abordar; cómo se ha desarrollado en la literatura relacionada; y los tipos de modelos aplicados usualmente en este contexto. Posteriormente, se describen los pasos seguidos en el desarrollo de la metodología, desde la adquisición de los datos hasta la evaluación de los modelos y la selección del mejor de ellos. Y finalmente, se identifican las variables que tienen un mayor impacto en la predicción del precio de los bienes inmuebles y se presentan los resultados obtenidos en cada etapa del desarrollo.

2. PLANTEAMIENTO DEL PROBLEMA

En las últimas décadas, el mercado de bienes inmobiliarios ha experimentado un rápido crecimiento, lo que finalmente se ha visto reflejado en el aumento de los precios de las viviendas [3]; estos cambios y la tendencia del mercado ha provocado un aumento en el interés en conocer el comportamiento de los precios anticipadamente, ya que los bienes raíces son una de las inversiones más críticas tanto para los compradores [4], como para los inversionistas que desarrollan proyectos para la venta, y adicionalmente para actores reguladores de este mercado. El mercado de bienes raíces tiene un importante rol en los sistemas económicos y sociales [5]; por lo que entender su comportamiento, puede guiar a la formulación de políticas de viviendas y programas de desarrollo económico, que ayuden a mejorar la equidad en la accesibilidad a la vivienda.

Debido a la variabilidad de la información en el mercado de bienes inmuebles, sobre los precios, las personas interesadas suelen afrontar pérdidas financieras [4]. Tener una estimación precisa del valor o del valor mediano de un inmueble es importante para todos los actores involucrados; para un inversionista dispuesto a diversificar su portafolio, esta información se considera de alto valor debido a las alternativas entre títulos de vivienda y otras posibles inversiones. Para los economistas e inversionistas, tener información de la predicción de los precios medianos de bienes inmuebles ayuda a tener una visión sobre las futuras tendencias de los precios, lo que permite optimizar el retorno de las inversiones y ayuda también a entender el mercado inmobiliario.

En el caso de los vendedores, conocer estos valores les permite evitar una sobrestimación o una subestimación del precio de venta del inmueble, lo que les evitaría perdidas innecesarias; y para los compradores es una ventaja reconocer entre un buen o un mal negocio para su inversión [6]. Las transacciones de compra y/o venta de viviendas puede ser una de las transacciones financieras más importantes para las familias [7], lo cual confirma le necesidad de tomar decisiones informadas, que les permita obtener el mayor beneficio en la compra o venta de bienes inmuebles.

Al ser un tema de alta conveniencia para grupos de interés de diferentes áreas se han llevado a cabo investigaciones relacionadas, buscando estimar de la manera más precisa para determinar el precio de los bienes inmuebles. Tradicionalmente, estas estimaciones se han centrado en el uso de variables relacionadas con los atributos estructurales de los inmuebles [8], posteriormente se fueron incluyendo variables de ubicación, de acceso a puntos de importancia, variables socioeconómicas del vecindario [6]; sin embargo, muchos de estos resultados son aplicables para ciudades no costeras, dejando de lado la evaluación de la importancia y efectos relacionados con atributos medioambientales de la región, e información sobre la distancia a la costa y sus posibles efectos . Todas estas variables analizadas de manera individual pueden no tener mucho impacto en la

estimación final del precio de un inmueble, pero al analizarlas de manera conjunta pueden ser variables significativas de la predicción buscada. [9]

Las investigaciones sobre la predicción de precios de bienes inmuebles suelen tener como limitación el acceso y la publicación de datos, ya que esta información no es de usual recolección. Es por esto que en algunas investigaciones se ha hecho uso de los precios medianos de las viviendas; como es el caso del trabajo del autor Vonlanthen, J , en el que hace uso de los precios medianos de las viviendas en Suiza para estudiar su relación con las tasas de interés hipotecarias y los rendimientos de bonos gubernamentales, analizando estos efectos por regiones [10] . Los precios medianos de bienes inmuebles han demostrado su efectividad en capturar información de los movimientos del mercado inmobiliario, en comparación con datos de ventas repetidas [11].

3. JUSTIFICACIÓN

Los inversionistas en el mercado inmobiliario suelen ver las zonas costeras como una atractiva inversión al ser un mercado con una alta demanda, liquidez y rendimientos duraderos; sin embargo, no se tiene mucha información de cómo estas inversiones y sus rendimientos puedan verse afectadas por factores de la región [12]. La zona Caribe es una de las regiones más urbanizadas a pesar de pertenecer a un grupo de regiones con ingresos bajos y medios. Esta urbanización no planificada distorsiona el equilibrio de la oferta y demanda de viviendas. [13]

De acuerdo con investigaciones que han evaluado los efectos del aumento del nivel del mar en el precio de bienes inmuebles en zonas costeras de Estados Unidos, se ha encontrado que el alto riesgo de aumento del nivel del mar puede llegar a tener un impacto estimado en una disminución del precio de bienes inmuebles en un 3.1% [14].

Los precios de las viviendas reflejan la calidad de vida y es un elemento clave para la productividad de una ciudad [15]. Es importante comprender cómo el entorno físico, la ubicación y atributos físicos de una propiedad impactan en el valor de una propiedad; ya que puede ayudar a entender el comportamiento de los precios en el mercado inmobiliario, como igualmente puede ser de utilidad para entidades gubernamentales a establecer políticas inteligentes en temas como accesibilidad, infraestructura, recuperación del valor de la tierra y facilitar restricciones financieras [16]

Los precios de las viviendas tienen importantes efectos en ámbitos socio económicos e incluso políticos, en los cuales pueden ayudar a proporcionar información que guíe la toma de decisiones informadas sobre programas de desarrollo económico y politicas de vivienda que ayude a la equidad en accesibilidad. Adicionalmente, al conocer esta información se pueden ayudar a economistas, analistas, inversionistas, entidades financieras e instituciones gubernamentales a preveer tendencias futuras y con esto tomar deciciones informadas sobre sus diferentes estrategias; lo cual es vital para la estabilidad económica de cualquier región.

Adicionando que la falta de información académica en esta industria y que el desarrollo de metodologías para la identificación de variables explicativas en los precios de los bienes inmuebles y de la predicción de este mismo, se ha centrado principalmente en ciudades no costeras; se propone llevar a cabo el planteamiento de una metodología que permita la estimación de los precios medianos del mercado inmobiliario en ciudades ubicadas en zonas costeras del Atlántico en EEUU, a partir de variables explicativas encontradas en una fase preliminar de búsqueda, extracción e ingeniería de características.

4. OBJETIVOS

4.1 OBJETIVO GENERAL

Realizar la predicción de precios del sector inmobiliario en zonas de la costa atlántica de EE. UU., mediante un enfoque de ML.

4.2 OBJETIVOS ESPECÍFICOS

- Extraer y realizar ingeniería de características a información acerca del sector inmobiliario.
- Determinar las variables explicativas y relevantes para la predicción de precios del sector inmobiliario en zonas costeras en Estados Unidos.
- Hacer uso de metodologías que permitan estimar la predicción buscada, para zonas donde no hay la información suficiente.
- Comparar técnicas de ensamble usando machine learning con métodos de Deep learning mediante el uso de un conjunto adecuado de métricas.

4 ESTADO DEL ARTE Y MARCO TEÓRICO

5.1 ESTADO DEL ARTE

La predicción de precios en el mercado inmobiliario ha sido de interés, no solo para inversionistas, compradores y vendedores, sino también para académicos que en las últimas décadas han desarrollado investigaciones en las que plantean y aplican metodologías para realizarla. Adicionalmente, a partir de sus resultados, se establecen diferentes variables explicativas y métodos para obtenerlas.

La predicción de precios del mercado inmobilario puede definirse como un modelo hedónico de precios, ya que cumple con la hipótesis de estos modelos que establece que los atributos relacionados deben analizarse como un conjunto; lo cual suele suceder en este tipo de problemas, ya que el valor de los bienes inmuebles se establece de acuerdo por el conjunto de sus atributos [16]. En la predicción de precios inmobiliarios, la dependencia entre los atributos y el precio de la propiedad es un factor que cambia con el tiempo [17]

Basándose en modelos hedónicso de precios, autores como Cao, calibra un modelo Ordinary Least Squares (OLS); sin embargo, al ser un modelo que pueda resultar con alto sesgo e inconsistencias en los resultados [18], emplea adicionalmente el modelo Geographically Weighted Regression (GWR) haciendo uso de la distancia euclidiana para medir la distancia espacial de las observaciones ,finalmente modifica el modelo GWR para basarse en una matriz tiempo y distancia [16]; permitiendo la no estacionariedad espacial y temporal [19].

La complejidad del problema de la predicción de los precios de bienes raíces y la no linealidad de los datos, causan que los métodos paramétricos no suelen tener los mejores resultados y por lo tanto se opta por usar métodos no paramétricos que suelen ser aplicados en problemas de modelos de valoración [20]. Otro factor por el cual se suelen aplicar los algoritmos de Machine Learning, es porque se obtienen mejores resultados en ajuste y precisión [6] . En el trabajo de Tchuente y Nyawa se consideran los modelos de Random Forest, Gradient Boosting, Adaboost, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) y Regresión Lineal; en este también hacen uso de variables relacionadas a características físicas de la propiedad, variables de accesibilidad, variables socioeconómicas del vecindario, variables medioambientales e integran variables de ubicación que suelen ser integradas en modelos hedónicos [6]. Otros autores como Louati también han implementado algoritmos de ML, debido a la mejoría de los resultados de precisión en muchas otras aplicaciones, concluyendo que el modelo de Random Forest performa mejor que los modelos de Decision Tree y Linear Regression [21].

Usar algoritmos no lineales de ML puede tener mejores resultados que métodos convencionales de regresión [22];sin embargo, en la mayoría de las

implementaciones de estos modelos en problemas de precios de bienes raíces se suele descuidar la no estacionariedad espacial y temporal [15]. La aplicación de modelos de Machine Learning como Linear Regression, Decision Tree, Random Forest y Gradient Boosting, mejoran los resultados en la predicción de precios de bienes raíces al incluir una variable de retraso espaciotemporal [15] ,como se establece en la investigación de Soltani ,en la cual se concluye que los modelos basados en técnicas de ensamble de Machine Learning como Random Forest o Gradient Boosting desempeñan mejor.

Según Kang, las investigaciones con interés en el mercado inmobiliario se han centrado en la estimación del precio y no en la tasa de apreciación de este, esta puede ser diferente y la apreciación del precio puede estar más relacionada con factores de apariencia física no solo de la propiedad, sino también de la zona en la que se encuentre ubicada, la decoración, entre otros [8]. Es por ello que este autor se plantea el uso de múltiples y diferentes fuentes de datos, combinando así características en datos estructurados y otras características extraídas a partir de imágenes de vista de la calle y fotografías de la propiedad y mediante el uso de técnicas de Deep Learning, y de la aplicación de modelos de Machine Learning para la predicción.

El uso de data estructurada junto con características extraídas de data no estructurada como las imágenes dan resultados más robustos y con mejor precisión [23]. Zhao propone un modelo hibrido entre Deep Learning y Machine Learning, en el cual usa inicialmente un modelo pre-entrenado de Convutional Neuroral Network (CNN) para evaluar características a partir de las imágenes y un modelo Multiplayer Perceptions (MLPs) para analizar los datos tabulares, y otro modelo CNN para la extracción de características de las imágenes; posteriormente crea una capa media conectada para analizar las características combinadas y finalmente la última capa se cambia por un modelo XGBoost de regresión para la predicción de los bienes raíces. Otros autores también han evaluado la efectividad de unir diferentes tipos de modelos, como Convutional Neuroral Network (CNN) con Random Forest, AdaBoost y con XGBoost, siendo esta última combinación la que mejor resultados se obtiene en términos de Mean Absolute Percentage Error (MAPE) [24] .

Chou et al. (2022) se desempeñan los modelos individuales y los modelos híbridos, usando Artificial Neuronal Networks (ANNs), Support Vector Machine (SVM), Classification and Regression Tree (CART) and Linear Regression (LR) como modelos individuales y modelos base para métodos de ensamble y modelos híbridos; obteniendo como resultado que el modelo hibrido Bagging-ANNs es mejor que los modelos base individuales o modelos de ensamble. Debido a su gran efectividad, los modelos híbridos entre Deep Learning y Machine Learning han estado adquiriendo más uso en diferentes problemas de regresiones o clasificaciones, usualmente en problemas que pueden ser mejorados al añadir características extraídas de imágenes relacionadas. Otro modelo Deep Learning usado como modelo individual y modelo base para modelos de ensamble en

aplicaciones de predicción de precios de bienes raíces es el modelo Multilayer Perceptron Neural Network (NEU), el cual en la investigación de Talaga et al. (2019) un MAPE superior.

La combinación de modelos de ML y DL ha probado ser efectiva y su uso se ha extendido por casos de uso de clasificación y regresión, específicamente en los casos en que es necesaria la extracción de características a partir de los anuncios en imágenes [4]. Chou y otros autores ,realizan pruebas de desempeño con modelos individuales y modelos híbridos, usando Artificial Neural Networks (ANNs), Support Vector Machine (SVM), Classification and Regression Tree (CART) y Linear Regression (LR) como modelos individuales y modelos base para métodos de ensamble y modelos híbridos; obteniendo como resultado que el modelo híbrido Bagging-ANNs es mejor que los modelos base individuales o modelos de ensamble. Adicionalmente, otro modelo Deep Learning usado como modelo individual y modelo base para modelos de ensamble en aplicaciones de predicción de precios de bienes raíces, es el modelo Multilayer Perceptron Neural Network (NEU),el cual presenta un desempeño mayor medido como MAPE [9].

Una de las limitaciones para analizar precios del mercado de bienes inmuebles es la dificultad para adquirir datos, por lo que las variables explicativas disminuyen y se dejan fuera del análisis y la predicción factores determinantes a la hora de predicción del precio. Algunas investigaciones han concluido que factores como la vista del vecindario y las cortas distancia a zonas de negocios están relacionados con un alto precio en los bienes inmuebles; sin embargo, se dejan por fuera factores del vecindario y ambientales que suelen reflejarse en el precio de las propiedades [18]. Los atributos estructurales como antigüedad de la construcción, cantidad de baños, habitaciones y número de pisos ocupados, suelen ser tradicionalmente asociados al precio de los bienes inmuebles; adicionalmente de características asociadas al vecindario tales como número de escuelas y paradas de transporte público cercanos a la propiedad; además de atributos ambientales como el nivel de ruido, calidad del aire, zonas verdes, entre otros [18]. Un factor adicional e importante que se puede considerar para el análisis de precios del mercado inmobiliario en zonas costeras es la cercanía con el mar, ya que se ha encontrado que las propiedades que tienen un mayor riesgo de inundación presentan un precio más bajo [14].

Una metodología usada para minimizar la limitación de datos es tomar propiedades similares; pero no es una solución que mejore resultados y no siempre se cuenta con dicha información. Una propuesta para solucionar este problema es unir diferentes zonas de la ciudad en zonas uniformes que reflejen características similares en el mercado inmobiliario [25]. En su propuesta, Lasota usa un modelo de regresión lineal y modelos de árboles, estableciendo como métrica de evaluación de los resultados las pruebas no paramétricas de Friedman y el de Wilcoxon para evaluar la significancia de los resultados.

La limitación de tener pocos datos es un inconveniente para la aplicación de los modelos que busquen la predicción de precios de bienes raíces en zonas de las cuales se tenga poca o ninguna información. El Transfer Learning es una metodología que permite usar un modelo que fue entrenado en un dominio distinto al de interés, para aplicarlo en el dominio de interés a pesar de tener pocos datos [26] y con esto poder estimar predicciones en situaciones donde se tiene poca información. Esta metodología suele ser usada en problemas de visión por computador; sin embargo, hay aplicaciones en problemas de predicción como el de predicir el movimiento del precio de las acciones a corto plazo, en el que se demuestra la efectividad del transfer learning [27].

5.2 MARCO TEÓRICO

Tomando como base las técnicas aplicadas por diversos autores, expuestas en la sección *ESTADO DEL ARTE*; para abordar problemas similares al que se plantea en el siguiente proyecto, y considerando los resultados más destacados, se tomarán los siguientes algoritmos como marco de referencia.

5.2.1 *K-nearest neighbors*

K-nearest neighbors es un algoritmo no paramétrico de Machine Learning que puede ser usado para problemas de clasificación y de regresión, basado en aproximaciones locales [28]. La idea base es tener un conjunto de Nobservaciones con variables x_i correspondientes a las características y una variable objetivo y_i; para el conjunto de atributos se calcula una métrica de distancia observando así los k vecinos más cercanos de x. En el caso de los problemas de regresión, la variable objetivo se aproxima de mejor manera usando un promedio ponderado. La ventaja de este método radica en su fácil implementación y bajo costo computacional, teniendo en cuenta que solo se requiere el parámetro k y la función de distancia; sin embargo, su desempeño en problemas de alta dimensionalidad es bajo [6]

5.2.2 *Random Forest*

El algoritmo Random Forest, es una combinación de predicciones de árboles de decisión de modo que cada uno de estos depende de un vector aleatorio independiente y tiene la misma distribución para todos los árboles de decisión [29]. Al tener un conjunto de N observaciones con variables de entrada x_i y una variable objetivo y_i , el primer paso consiste en seleccionar una muestra aleatoria del conjunto total, con reemplazo; por cada una se forma un árbol de decisión T_1, en el cual para cada nodo se selecciona aleatoriamente p variables de entrada y se usan para dividir el nodo basado en la reducción de varianza. Lo anterior se repite m veces y se obtienen T_m arboles de decisión; adicionalmente, la predicción de y se obtiene de promediar las predicciones de cada árbol. Este algoritmo tiene mejor

precisión y reduce problemas de varianza; sin embargo, tiene un alto costo computacional. [6]

5.2.3 *Gradient Boosting*

Gradient Boosting es una técnica de boosting que permite la optimización con otras funciones de pérdida diferenciables [6]; siendo la principal idea del boosting, agregar nuevos modelos conjuntos secuencialmente. En el proceso de entrenamiento se ajustan consecutivamente nuevos modelos para proporcionar una estimación más precisa de la variable objetivo. La principal idea es construir los nuevos algoritmos bases de aprendizaje mayormente correlacionados con el gradiente negativo de la función de pérdida, la cual suele ser el error cuadrático. [30]

5.2.4 *AdaBoost*

Adaboost toma los algoritmos bases de aprendizaje más débiles como un conjunto, convirtiéndolo en un solo algoritmo de aprendizaje más fuerte al tener mayor estabilidad que un sólo árbol de decisión complejo [6]. Una de las principales ideas es mantener una distribución de pesos en el conjunto de entrenamiento, los cuales inicialmente son los mismos; sin embargo, en cada iteración los pesos de las predicciones incorrectas se aumentan, de manera que el algoritmo base débil se deba esforzar más en el aprendizaje de los puntos más difíciles del conjunto de aprendizaje. [31]

5.2.5 Extreme Gradient Boosting (XGBoost)

En contraste con los árboles de decisión, XGBoost aprende de los datos predecesores y también agrega puntuaciones en las ramas correspondientes para reducir los errores del árbol anterior [23]. Este algoritmo es una técnica de boosting que domina las predicciones de un conjunto de algoritmos base débiles para desarrollar uno fuerte a través de estrategias de entrenamiento aditivo. XGBoost simplifica la función objetivo al permitir combinar términos predictivos y de regularización. El entrenamiento aditivo consta en que el primer algoritmo base de aprendizaje se ajusta a todo el conjunto de datos de entrada y luego se ajusta un segundo algoritmo base a estos residuos del algoritmo de aprendizaje débil, este ajuste se repite hasta cumplir un criterio de parada; la predicción final es la suma de la predicción de cada algoritmo de aprendizaje. [32]

5.2.6 *Artificial Neural Networks (ANN)*

En una comparación con las redes neuronales biológicas, los algoritmos de redes neuronales están compuestos por neuronas artificiales llamadas nodos y estos están conectados entre sí a través de bordes, los cuales transmiten señales que están asociadas a números y son conexiones entre la salida de un nodo y la entrada de otro [6]. El algoritmo ANN tiene tres diferentes tipos de capas: entrada, oculta y

salida. Cada capa consiste en un número de neuronas o nodos, los cuales están conectados con los nodos de la capa siguiente; las conexiones o bordes, representan un peso que se ve reflejado en el ajuste [33]. Para calcular la salida de un nodo especifico, las señales entrantes se combinan con los pesos de todos los bordes de la entrada y el sesgo del nodo se ajusta con una función de transferencia; este proceso se aplica para todos los nodos hasta obtener la estimación de la última salida. Este algoritmo presenta mayores ventajas en datos no lineales. [6]

6 METODOLOGÍA

El presente trabajo se lleva a cabo bajo una adaptación de la metodología Cross-Industry Standard Process for Data Mining propuesta por Wirth & Hipp (2000), siguiendo las fases descritas a continuación.

6.1 *ENTENDIMIENTO DEL NEGOCIO:*

En esta fase se realizó una revisión de la literatura sobre las aplicaciones de modelos de regresión en machine learning para la predicción de precios de bienes inmobiliarios, enfocando el interés sobre los algoritmos utilizados y las variables usadas en dichas aplicaciones. Esta fase impacta en la decisión de cuales modelos tener en cuenta para la fase de entrenamiento y evaluación, adicional de cuales variables se consideran importantes buscar en los datos disponibles.

6.2 *ENTENDIMIENTO, PREPARACIÓN Y ANÁLISIS DE DATOS:*

Se obtienen los datos de acuerdo con la disponibilidad que se puede encontrar en la página oficial del Censo de los Estados Unidos, reuniendo información acerca de los precios, pero también de variables que permiten hacer un análisis para determinar sus posibles relaciones e impactos. Esto se lleva a cabo haciendo uso de la API The American Community Survey (ACS) que contiene información sobre características de vivienda, económicas y demográficas, entre los años 2009 a 2022. Posteriormente, se evalúa la calidad de los datos, se aplican técnicas de análisis descriptivo que permitan comprender el estado de los datos obtenidos y extraer información que estos contengan. Con el fin de mejorar alguna situación presente en la calidad de los datos, se pueden seleccionar un subconjunto de las variables a usar a partir de diferentes métodos a aplicar, según los resultados de los análisis realizados.

6.3 *MODELACIÓN:*

Se entrenan modelos con la aplicación de los algoritmos descritos en la sección *MARCO TEÓRICO* como Multiple Linear Regression, K-nearest neighbors, Random Forest, Gradient Boosting, AdaBoost, XGBoost, Artificial Neuronal Networks. Se realizan iteraciones con enfoques y aplicación de técnicas diferentes con el objetivo de tener un amplio grupo de resultados que permitan hacer una comparación y evaluación más completa.

6.4 *EVALUACIÓN:*

Una vez los modelos están entrenados y con hiperparámetros optimizados, se procede con la evaluación, tomando como referencia las métricas usadas en la literatura en problemas de regresión. A continuación, se definen las métricas a usar.

- ***Root Mean squared error (RMSE):*** Mide que tan dispersos están los residuos, es decir, los errores de la predicción. Es una métrica que proporciona el error de predicción en unidades de la variable de interés.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Ecuación 1. Root Mean Squared Error (RMSE)

- ***R2:*** Representa la proporción de varianza de la variable objetivo y que es explicada por las variables de entrada x_i

$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

Ecuación 2. R2

7 DESARROLLO DEL TRABAJO

El desarrollo de este trabajo se enfoca en abordar el problema descrito en la sección *PLANTEAMIENTO DEL PROBLEMA*, comenzando por la obtención de un conjunto de datos que contiene información sobre variables inmobiliarias y sus precios en una zona costera seleccionada. Luego, se evalúa el estado de los datos, en el que se verifica su calidad, la cantidad de información disponible, la diversidad y tipos de variables presentes. Se aplican técnicas de análisis descriptivo, como la exploración de distribuciones, cantidades, promedios y otros análisis, que permiten comprender el estado de los datos recopilados.

Una vez se lleva a cabo el análisis descriptivo y de calidad de los datos, se aplican métodos de selección de características que permiten eliminar variables irrelevantes o redudantes, dejando asi las de mayor impacto y ayudando a la interpretabilidad del modelo. Después de definir las variables a utilizar, la cantidad de datos y sus transformaciones, se aplicaron los modelos descritos en el *MARCO TEÓRICO* con el objetivo de predecir los precios de los inmuebles en las zonas costeras incluidas en el conjunto de datos. Finalmente, se evaluaron los modelos y se seleccionó el mejor basándose en métricas como $RMSE$ y R^2.

El desarrollo del presente trabajo, se puede observar en el repositorio público de github : prediccion-real-estate

7.1 CONJUNTO DE DATOS SELECCIONADOS

La muestra de datos elegida para el presente trabajo comprende variables inmobiliarias, socioeconómicas, microeconómicas y demográficas para los condados de 6 estados de la Costa Este de los Estados Unidos: Florida, Carolina del Sur, Carolina del Norte, Virginia, Nueva York y Nueva Jersey. La muestra cubre el período de tiempo de 11 años, entre 2011 y 2022; y fue seleccionada en función de la disponibilidad de los datos transversales más recientes que se pueden encontrar en la Oficina del Censo de los Estados Unidos.

Se proporciona una lista de las 23 variables incluidas en el conjunto de datos en la Tabla 1.

Nombre variable	***Definición***
State	Estado
County	Condado
Housing_units	Total de viviendas
Median_rooms	Mediana de cantidad de habitaciones por vivienda

Median_Price	Precio mediano de vivienda. Variable objetivo
Total_population	Total de la población
Median_age	Mediana de la edad de la población
Vacant_housing_units	Total de viviendas desocupadas
Owner_occupied	Total de viviendas ocupadas por el propietario
Renter_occupied	Total de viviendas ocupadas por inquilinos
Total_Household	Total de hogares familiares
Median_Household_income	Mediana del ingreso por hogar
Median_Family_income	Mediana del ingreso por familia
PerCapita_income	Ingreso per capita, ingreso promedio por persona
Nonfamily_households	Total de hogares no familiares
Median_nonfamily_income	Mediana del ingreso el hogares no familiares
Median_Gross_Rent	Mediana de la renta bruta
Gini_Index	Medida de desigualdad en los ingresos de una población
Poverty_Status	Población total para la determinación de pobreza
Unemployment_ 16YearsAndOver	Desempleo en personas con edad de 16 años o mayores
# State	Código del estado
# County	Código del condado

Tabla 1. Variables del conjunto de datos original

Se selecciona la variable ***Median_Price*** como la variable objetivo, dado que representa el precio de los bienes inmuebles y es el dato que se busca predecir en este proyecto. Es relevante considerar que los datos están agrupados a nivel de estado y año; no obstante, los análisis y las predicciones se realizarán a nivel general debido a consideraciones de la cantidad de datos disponibles. El conjunto de datos recopilados consta un total de 5151 datos y 23 variables, en la Tabla 2 se presenta la distribución de cantidad de datos por estado.

Estado	***Cantidad de datos***
Florida (FL)	804
Carolina del Norte (NC)	1199
New Jersey (NJ)	252
New York (NY)	744
Carolina del Sur (SC)	552
Virginia (VA)	1599

Tabla 2. Cantidad de datos por estado

7.2 ENTENDIMIENTO Y ANÁLISIS DE LOS DATOS

Se emplean técnicas de análisis descriptivo para comprender la calidad de los datos, la completitud, conformidad y duplicidad de los datos. Adicional se realizan algunos tipos de gráficos para visualizar la distribución de los datos de variables específicas.

Para eliminar los datos atípicos en el conjunto de datos usado en este proyecto, se aplica una técnica basada en el uso de quantiles y el rango intercuartílico (IQR). Los datos se agrupan por la variable *State* para asegurar que el análisis de outliers se realice dentro de cada grupo de estado, posteriormente se calculan los cuantiles Q1 y Q3 de *Median_Price* para cada grupo. Luego se definen los limites, siendo el limite inferior $Q1 - 1.5 * IQR$ y el limite superior $Q3 + 1.5 * IQR$. Finalmente se eliminan los registros que tenían valores de *Median_Price* fuera de estos límites.

La presencia de outliers, o datos atípicos, puede causar sesgos en los análisis e introducir ruido, llevando a problemas de sobreajuste y reduciendo la capacidad de generalización del modelo, afectando la calidad de la predicción de la variable objetivo. La eliminación de estos datos atípicos ayuda a mejorar la calidad y confiabilidad de los análisis a realizar y los modelos predictivos a desarrollar.

Se escalan los datos con el objetivo de normalizar el rango de las variables continuas, llevándolas a un rango común. Esto ayuda a que los análisis a realizar de componentes principales, correlaciones, multicolinealidad y la selección de características, no se vean influenciados de manera desproporcionada por las variables de mayor rango; adicionalmente, ayuda a que los modelos tengan un mejor desempeño y generen predicciones más precisas.

También lleva a cabo el Análisis de Componentes Principales (PCA) ,para reducir la dimensionalidad del data set compuesto por 23 variables; con el objetivo de facilitar la visualización y el análisis interpretativo al proyectar los datos en un espacio de menor dimensión, conservando la mayor parte de la variabilidad original. Además, el PCA ayuda a identificar patrones y relaciones subyacentes, eliminando redundancias y colinealidad entre las variables, además de mejorar la interpretabilidad de los resultados. Teniendo en cuenta los hallazgos del análisis de componentes principales y su influencia en las variables a través del PCA, se procede a realizar un análisis de correlación para el conjunto de datos original sin datos atípicos, ya que estos pueden distorcionar las relaciones entre las variables.

La correlación de Pearson es usualmente usada para medir la correlación entre dos variables cuando la variable objetivo a medir es de tipo continuo [34], como es el caso de los datos a analizar en el presente proyecto. El coeficiente de correlación Spearman se puede usar como alternativa al coeficiente de correlación Pearson ya

que puede ser menos sensible a datos atípicos al usar una clasificación de rangos de los datos originales, evaluando una relación monotónica pero no necesariamente lineal [35] como sucede con el coeficiente de Pearson.

Con el propósito de explorar y comprender mejor cómo se relacionan las variables del conjunto de datos utilizado, se realiza un análisis de dependencias lineales para identificar si hay variables que influyan directamente en otras o si alguna variable puede ser expresada en términos de otra. En caso de existir tales dependencias, se procederá a eliminar las variables redundantes del conjunto de datos utilizado para entrenar el modelo de predicción. Esto se hace debido a que la presencia de dependencias lineales puede ocasionar problemas de sobreajuste u overfitting, lo que resulta en predicciones poco precisas cuando el modelo es aplicado a datos no observados. El sobreajuste reduce la capacidad de generalización del modelo, por lo que es crucial abordar esta situación.

7.3 SELECCIÓN DE CARACTERÍSTICAS

Para abordar los desafíos de la multicolinealidad que se pueden presentar ,es esencial realizar una selección de características utilizando varios métodos; estos incluyen correlaciones tradicionales, índice de correlación múltiple, que permiten identificar y eliminar variables redundantes. Adicionalmente, técnicas como la selección forward y backward; también la regresión Lasso y Random Forest que proporcionan enfoques robustos para la selección de características. A cada conjunto resultante se le evaluará la dependencia lineal y la multicolinealidad para determinar si se mejora con respecto al conjunto de datos originales. Además, con cada conjunto se entrenará una regresión lineal múltiple y se evaluará utilizando el RMSE y el R^2 para identificar cuál conjunto resultante podría performar mejor en un posible modelo de regresión. Estas técnicas ayudan a reducir la dimensionalidad del modelo, minimizar la multicolinealidad y mejorar la interpretabilidad y confiabilidad en los resultados obtenidos de los modelos predictivos.

Para evaluar los conjuntos de datos seleccionados por cada técnica de selección de características mencionadas, se utiliza una partición de datos en un conjunto de entrenamiento y una prueba con una proporción del 75% y 25%, respectivamente. Esta partición tiene en cuenta el orden cronológico del conjunto de datos, de acuerdo con el valor de la variable *Year*; se ordenan los datos por dicha variable y para el conjunto de prueba se toma el 25% final del conjunto de datos original, de esta manera el conjunto de prueba tiene en cuenta los datos de los últimos años.

Es importante tener en cuenta que, en esta etapa del análisis, no se utiliza un conjunto de validación porque el objetivo no es construir el modelo de predicción final, sino evaluar cómo performaría el conjunto de datos seleccionado en una posible predicción. Esta metodología permite obtener una primera estimación del rendimiento del modelo y entender la viabilidad de las variables seleccionadas antes de proceder a una validación más rigurosa y a la construcción del modelo definitivo.

7.4 MODELACIÓN Y OPTIMIZACIÓN DE HIPERPARÁMETROS

Para cada uno de los algoritmos mencionados en el *MARCO TEÓRICO*, se lleva a cabo el entrenamiento y la evaluación de modelos de predicción, donde la variable objetivo es *Median_Price*. Esto se realiza usando el conjunto de variables seleccionadas que haya dado los mejores resultados en la evaluación descrita anteriormente.

El conjunto de datos seleccionado, evaluado y analizado en el presente proyecto proviene del Censo oficial de Estados Unidos; al ser un censo, la información se encuentra de manera anual sin registros de fechas especificas. El conjunto de datos a usar en la predicción de la variable objetivo, luego eliminar datos atípicos y hacer una selección de variables, consta solo de variables continuas que abarcan registros desde 2011 a 2022; por lo que se decide tomar los registros del último año (2022) como conjunto de validación, y los datos de 2011 a 2021 se usan para el entrenamiento y la evaluación de los modelos de regresión definidos en el *MARCO TEÓRICO*.

Adicionalmente, al conjunto de variables seleccionadas, se incluyen variables State y Year. Esto se realiza con el objetivo de controlar efectos de tendencia de acuerdo con el tiempo y las áreas de territorio ocupadas por los diferentes estados. En el caso de la variable Year, al ser una variable que implica un orden progresivo, se realiza una codificación usando LabelEncoding. Para la variable State se usan variables dummies derivadas de esta, ya que al no tener un orden natural, esta codificación permite a los modelos evaluar cada modelo de mandera independiente; en este proceso se elimina la primera variable dummy para evitar multicolinealidad, ya que la inclusión de todas las variables dummy junto con la constante del modelo puede causar problemas de dependencia lineal.

Para realizar la partición del conjunto de datos para entrenamiento y evaluación, se tiene en cuenta el orden cronológico por la variable *Year.* Nuevamente, tras mantener el ordenamiento, se toma el 25% final del conjunto de datos; teniendo de esta manera un conjunto de entrenamiento con una proporción de 75% correspondiente a los primeros años y un conjunto de prueba con el 25% de los datos, cuya información corresponde a los últimos años. Dado que las variables a usar son continuas, no se requiere realizar un muestreo estratificado; ya que no hay categorías específicas que necesiten estar representadas proporcionalmente en los conjuntos de entrenamiento y prueba, por lo que no se pone en riesgo la representatividad de los datos al utilizar esta técnica.

Inicialmente se decide tomar los modelos expuestos en el *marco teórico* de este documento, con sus parámetros por defecto; esto se realiza con el objetivo de evaluar el rendimiento base de cada modelo sin la influencia de ajustes específicos de hiperparámetros, proporcionando una referencia en la comparación de los

resultados. Posteriormente, se buscan los mejores hiperparámetros mediante técnicas comúnmente usadas, como GridSearch y RandomizedSearch.

RandomizedSearch selecciona combinaciones de hiperparámetros de manera aleatoria dentro de un espacio de búsqueda definido, evaluando solo un número fijo de configuraciones, por lo que puede ser más eficiente en términos de tiempo computacional [36]. Se establecen los siguientes rangos de hiperparámetros para cada modelo

- **Random Forest Regressor:**
  - n_estimators: [10,100]
  - max_depth: [5,10]
  - max_features: [1,13]
  - min_samples_split: [2,11]
  - min_samples_leaf: [1,11]
  - criterion: [mse,mae]
- **Lasso Regression:**
  - alpha: [1, 100]
- **KNeighbors Regressor:**
  - n_neighbors: [1,100]
  - weights: [uniform, distance]
  - p: [1,2]
- **Gradient Boosting Regressor:**
  - n_estimators: [10, 200]
  - max_depth: [1,10]
  - learning_rate: [0.01, 1]
  - subsample: [0.5, 1]
- **AdaBoost Regressor:**
  - n_estimators: [10, 200]
  - learning_rate: [0.01, 1]
- **XGBoost Regressor:**
  - n_estimators: [10, 200]
  - max_depth: [1, 10]
  - learning_rate: [0.01, 0.1]
  - subsample: [0.5, 1]
  - colsample_bytree: [0.5, 1]
- **Keras Regressor (ANN):**
  - optimizer: [adam, rmsprop]
  - activation: [relu, tanh]
  - loss: [mse, mae]
  - batch_size: [16, 32]
  - neurons: [16, 32]
  - epochs: [20, 50]
  - patience: [2, 5]

GridSearch es un enfoque en el que se realiza una búsqueda exhaustiva de los mejores hiperparámetros, usando un conjunto específico de estos predefinidos, y evalúa todas las combinaciones posibles para encontrar la mejor combinación [36]. Se evalúan los siguientes hiperparámetros para cada modelo, con la optimización GridSearchCV.

- **Random Forest Regressor:**
  - n_estimators: 10, 20, 30, 40, 50
  - max_depth: 15, 20, 30, 50
  - max_features: 3, 5, 7, 10
  - min_samples_split: 3, 5, 7, 10
  - min_samples_leaf: 1, 3, 5
  - criterion: mse,mae
- **Lasso Regression:**
  - alpha: 0.05, 0.1, 0.05, 1, 5, 10
- **KNeighbors Regressor:**
  - n_neighbors: 3, 5, 7, 10, 15, 20
  - weights: uniform, distance
  - p: 1, 2
- **Gradient Boosting Regressor:**
  - n_estimators: 50, 100, 150, 200
  - max_depth: 3, 5, 7, 10
  - learning_rate: 0.01, 0.1, 0.2
  - subsample: 0.5, 1
- **AdaBoost Regressor:**
  - n_estimators: 50, 100, 150, 200
  - learning_rate: 0.01, 0.1, 0.2
- **XGBoost Regressor:**
  - n_estimators: 50, 100, 150, 200
  - max_depth: 3, 5, 7, 10
  - learning_rate: 0.01, 0.1, 0.2
  - subsample: 0.5, 1
  - colsample_bytree: 0.5, 1
- **Keras Regressor (ANN):**
  - optimizer: adam, rmsprop
  - activation: relu, tanh
  - loss: mse, mae
  - batch_size: 16, 23, 32
  - neurons: 16, 23, 32
  - epochs: 20, 35, 50
  - patience: 2, 3, 5

Adicionalmente, se emplea la estrategia de validación cruzada con series temporales (TimeSeriesSplit) para asegurar una evaluación más confiable en datos secuenciales. En este caso, el conjunto de entrenamiento se divide en 5 partes o splits. Cada split se usa de manera que los datos de entrenamiento incluyen todos los puntos de tiempo previos y el split restante se usa como conjunto de prueba, repitiendo este proceso 5 veces. En cada iteración se calcula el error cuadrático medio (MSE), y se promedian los resultados para evaluar el rendimiento general del modelo y seleccionar los mejores hiperparámetros.

Una vez que se seleccionan los mejores hiperparámetros para cada modelo, basándose en los resultados de la validación cruzada, el modelo se entrena nuevamente con el conjunto original de entrenamiento y se evalúan las métricas $RMSE$ y R^2 el conjunto original de entrenamiento y prueba, con el objetivo de evaluar que no exista sobreajuste. Adicionalmente, se evalúan dichas métricas con el conjunto de validación definido anteriormente, con el objetivo de evaluar los modelos en datos más recientes y confirmar que generelicen bien.

7.5 ELECCIÓN MEJOR MODELO

La selección del mejor modelo se basa en el menor valor de $RMSE$, también se tiene en cuenta el valor R^2 para evaluar la capacidad explicativa del modelo con respecto a la variable objetivo. Adicionalmente, se tiene en cuenta las diferencias de las métricas en los conjuntos de entrenamiento, prueba y valdiación, con el objetivo de verificar que no exista sobreajuste en el modelo. Finalmente, se obtiene el nivel de importancia de cada variable en el mejor modelo y se analizan las gráficas de residuales y de ajuste de las predicciones vs los valores reales.

8 RESULTADOS

8.1 ENTENDIMIENTO Y ANÁLISIS DE LOS DATOS

8.1.1 Análisis descriptivo

El conjunto de datos recopilados consta de un total de 5151 registros y 23 variables; tiene un registro nulo en la variable *Median_nonfamily_income*, el cual corresponde al condado Camden del estado de Carolina del Norte en el año 2015. Al ser el único registro nulo en una sola variable, se toma la decisión de eliminar dicho registro, ya que representa el 0.01% del total de los datos y no tiene un impacto significativamente alto. El resto de las variables del conjunto de datos tiene una completitud del 100% y no se encuentra duplicidad en ningún registro.

En la Tabla 3 se detalla el tipo de dato para cada variable, donde se encuentran 16 variables con datos tipo entero (integer), 4 variables con datos tipo flotante (float) y 2 variables con datos tipo cadena de caracteres (string).

Nombre variable	***Tipo de dato***
State	String
County	String
Housing_units	Integer
Median_rooms	Float
Median_Price	Integer
Total_population	Integer
Median_age	Float
Vacant_housing_units	Integer
Owner_occupied	Integer
Renter_occupied	Integer
Total_Household	Integer
Median_Household_income	Integer
Median_Family_income	Integer
PerCapita_income	Integer
Nonfamily_households	Integer
Median_nonfamily_income	Float
Median_Gross_Rent	Integer
Gini_Index	Float
Poverty_Status	Integer
Unemployment_ 16YearsAndOver	Integer
# State	Integer

# County	Integer

Tabla 3. Tipos de datos de las variables

Se tienen registros de los años 2011 a 2022 para las variables seleccionadas de acuerdo con el contexto del planteamiento del presente proyecto. Esta información se distribuye entre 6 estados, con 368 condados. Es importante tener en cuenta que cada condado tiene un registro por año, por lo que la cantidad de datos para cada estado está relacionada con la cantidad de condados. Sin embargo, se encuentra que para el condado Camden en el estado de Carolina del Norte no hay información para el año 2015, esto se debe a la eliminación del registro por la variable *Median_nonfamily_income* con valor nulo; adicional, el condado Bedford City del estado de Virginia solo contiene información para los años 2011 a 2013.

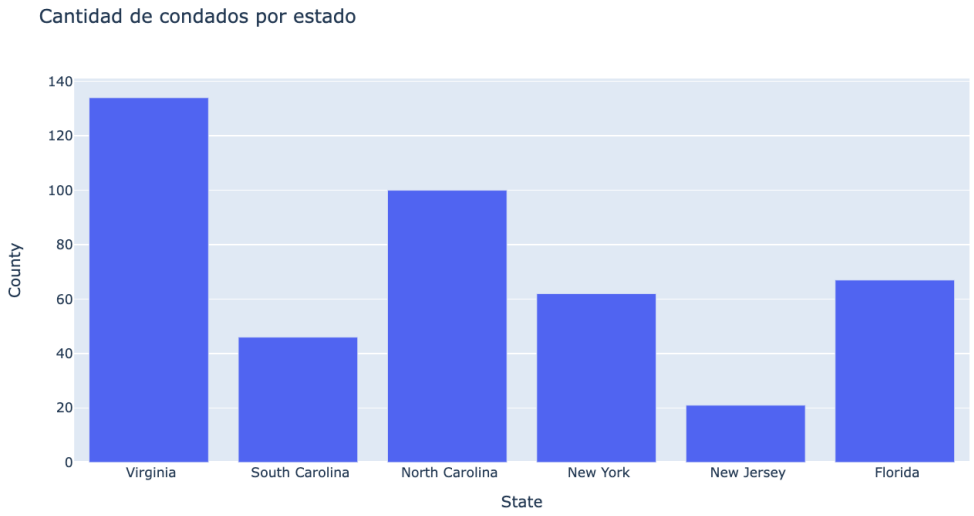

Figura 1. Cantidad de condados por estado

En la Figura 1 se muestra la distribución de la cantidad de condados por estado, en la que se observa que los estados con mayor cantidad de condados son Virginia y Carolina del Norte, y el estado con mejor cantidad de condados es New Jersey; los cuales concuerdan con los estados con mayor y menor cantidad de registros, respectivamente, lo que se puede confirmar con la información de la Tabla 2.

La variable Median_Price tiene una completitud del 100% de los registros del conjunto de datos; ya que se establece como la variable objetivo, es fundamental analizar su distribución y determinar si existe dispersión. Se aclara que los valores están expresados en dólares.

En la Figura 2 se observa que el valor del rango intercuartílico (IQR) es de 100.000 USD, ya que la caja del boxplot esta en el rango de valores del cuartil 1 (Q1) 150.000 USD hasta el valor del cuartil 3 (Q3) 250.000 USD. Adicionalmente, los bigotes se extienden desde alrededor de 50,000 USD hasta aproximadamente 400,000 USD; sin embargo, hay varios valores atípicos por encima del bigote superior, algunos

incluso superando los 1,000,000 USD. Esto indica que, aunque la mayoría de los precios medianos se encuentran dentro de un rango moderado, existen algunos casos con precios significativamente más altos. También se observa que no hay una distribución simétrica en la variable objetivo, ya que los datos están más distribuidos hacia el bigote superior, indicando que los valores de esta variable tienden a ser altos. Adicionalmente, la mediana no está centrada, sino que está cerca de los 200,000 USD.

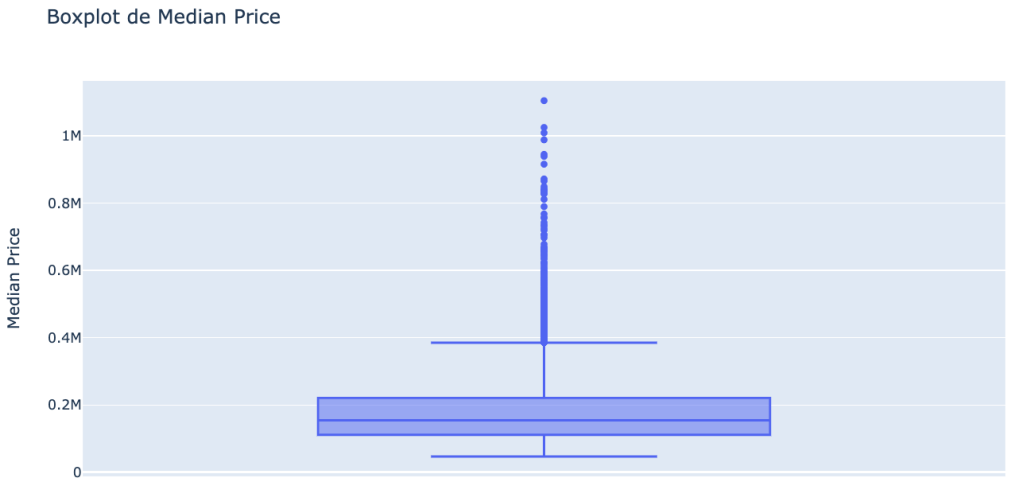

Figura 2. Boxplot Median_Price

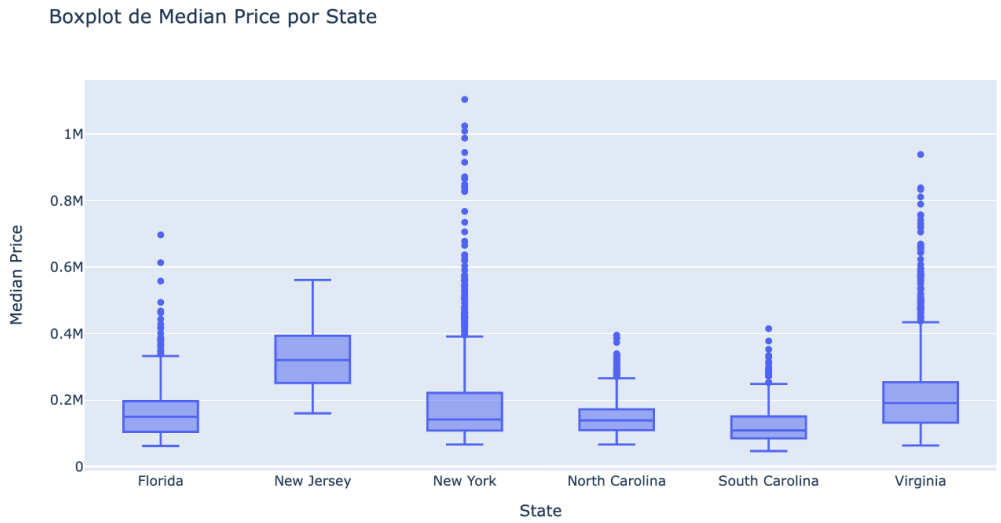

Figura 3. Boxplot Median_Price por State

En la Figura 3, se observa que en todos los estados la distribución de los datos no es simétrica al comparar visualmente la longitud del bigote inferior y superior en el gráfico de boxplot. Para el estado de New Jersey (NJ) se aprecia la presencia de pocos datos atípicos; su caja intercuartil cuya línea mediana se encuentra en el

centro en el segundo cuartil (Q2) no se encuentra más cercana a alguno de los límites, como sucede en otros estados observados dentro de la figura.

Para los estados de Florida (FL), New York (NY), Carolina del Norte (NC), Carolina del Sur (SC) y Virginia (VA) se visualizan datos atípicos por encima del rango intercuartil; este comportamiento se observa con mayor volumen en los estados NY y VA. Además, se puede notar que la línea mediana de la caja intercuartil no se encuentra centrada en los casos de NY y SC. Los estados de CN, FL y VA tienen un comportamiento similar al tener datos atípicos, no tener una distribución simétrica de sus bigotes y contar con una línea mediana más centrada que en otros estados, como NY y SC. Los estados de VA y NJ destacan por tener los precios medianos más altos alrededor de 300.00 USD y 400.000 USD, respectivamente; por el lado de FL, NC y SC se observan medianas más bajas, con valores menores a los 200.000 USD; estas diferencias destacan la variabilidad en los precios medianos de las viviendas entre los distintos estados.

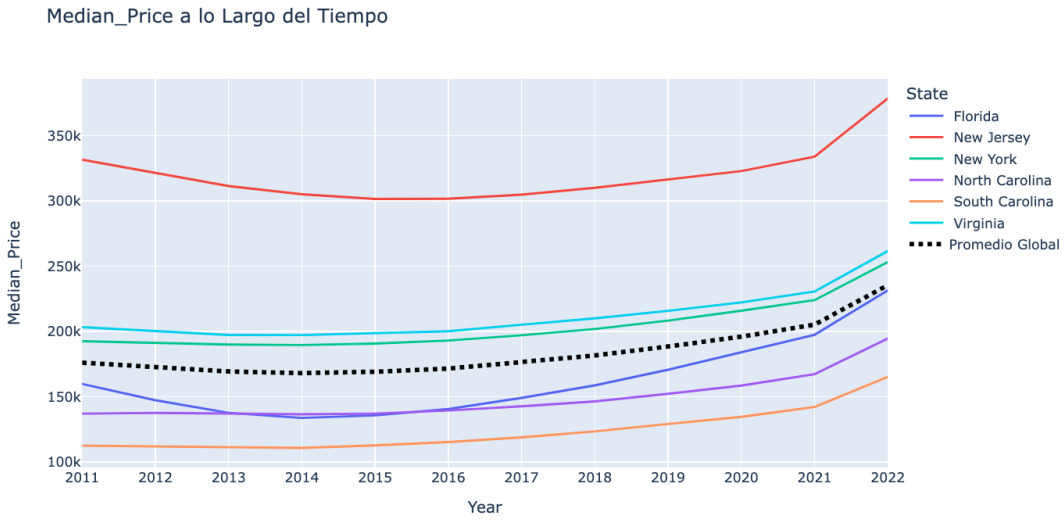

Figura 4. Evolución Median_Price a lo largo del tiempo, por estado y en promedio

En la Figura 4, se observa la evolución del *Median_Price* por estado a lo largo del tiempo, junto con una línea punteada negra que representa el promedio del conjunto total. La gráfica revela que, aunque todos los estados muestran una tendencia general al alza en los precios medianos, existen diferencias notables en los niveles de precios y en la tasa de incremento. New Jersey, New York y Virginia continúan destacándose con precios medianos más altos a lo largo del tiempo, mientras que estados como South Carolina y North Carolina mantienen precios más bajos. La línea del promedio total destaca claramente en la gráfica, mostrando una tendencia ascendente más moderada en comparación con algunos estados individuales.

State	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
Florida	$162.201,62	$76.643,79	$62.400	$104.800	$150.250	$196.900	$696.900
New Jersey	$320.015,48	$90.402,96	$160.500	$252.175	$320.450	$393.575	$561.500
New York	$203.987,37	$161.055,03	$66.800	$108.700	$141.950	$221.725	$1.104.000
North Carolina	$148.914,26	$54.357,19	$66.800	$109.900	$139.300	$172.450	$395.600
South Carolina	$124.029,17	$55.066,45	$47.100	$85.175	$109.150	$150.950	$415.000
Virginia	$211.889,99	$114.798,32	$63.900	$132.450	$191.600	$254.000	$938.500

Tabla 4. Distribución Median_Price por Estado

En la Tabla 4 se presenta la distribución de los datos de la variable objetivo *Median_Price* por estado, lo que permite observar y comparar el comportamiento de dicha variable. El estado de NJ tiene el mayor promedio de *Median_Price*, seguido por el estado de VA. NJ también tiene el valor mínimo y el valor mediano más alto; sin embargo, es importante destacar que el estado de NJ no cuenta con uno de los valores máximos de *Median_Price* más elevados. Por otro lado, NY y VA si tienen los valores máximos más altos; a esto se le suma el hecho de que también son los estados con la mayor variabilidad en el precio, como se observa en el valor de la desviación estándar (std) y en el gráfico de la Figura 2.

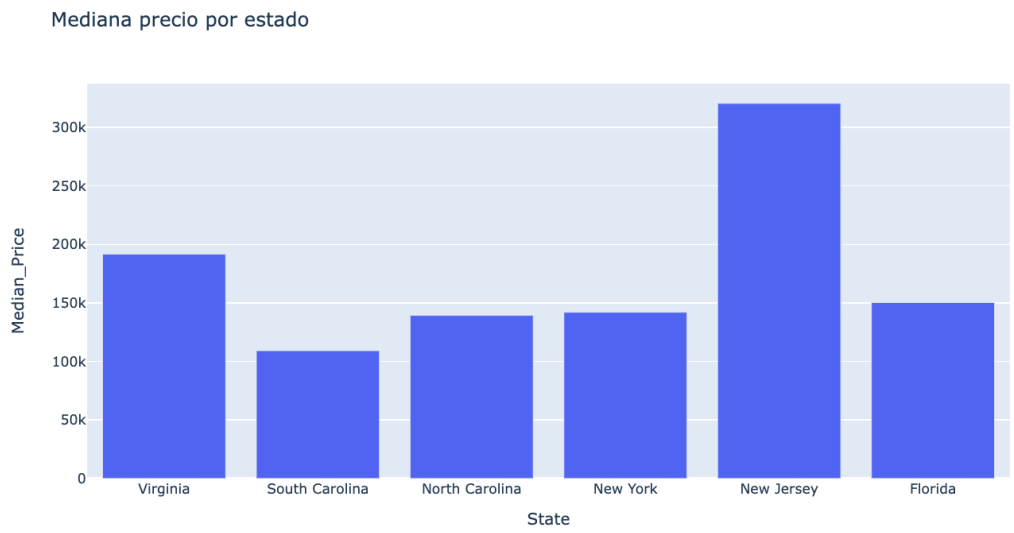

Figura 5. Distribución de las medianas de los precios medios por los estados analizados

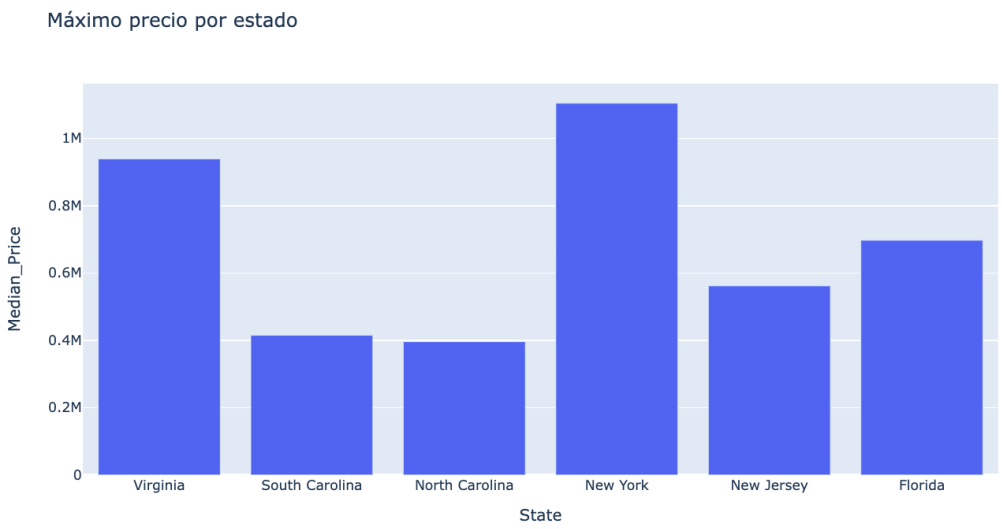

Figura 6. Distribución de los valores máximos de los precios medios por los estados analizados

Con las figuras anteriores, se confirma que los estados con mayores *Median_Price* son VA, NJ y NY. Aunque NY no tienen el valor más alto en promedio, sí tiene el valor máximo lo cual puede explicarse debido a la alta dispersión que presenta la variable objetivo en este estado, como se observa en los valores de la Tabla 4 y en el gráfico de boxplot en la Figura 3.

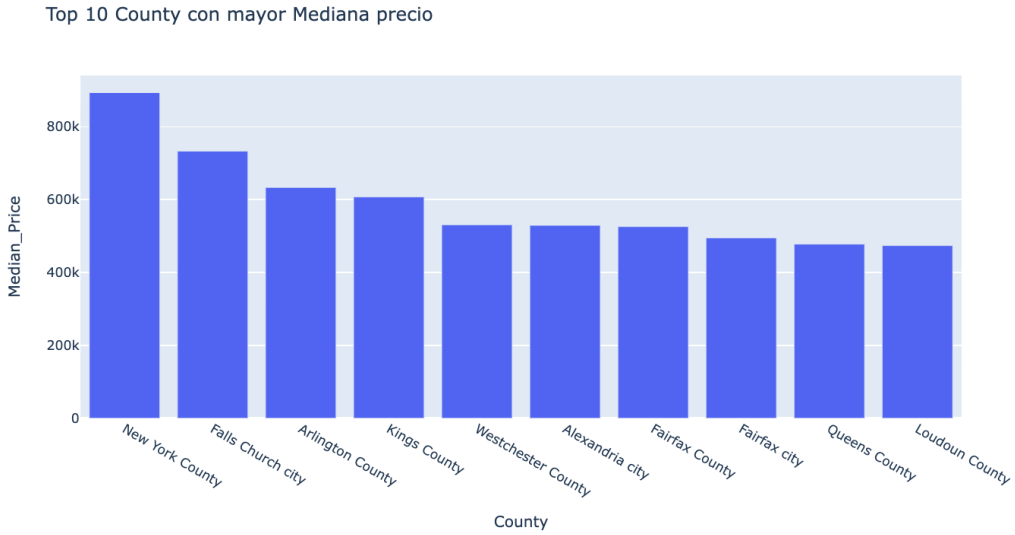

Figura 7. Top 10 condados con mayor MedianPrice

County	***Estado***
New York County	NY
Falls Church city	VA
Arlington County	VA

Kings County	NY
Westchester County	NY
Alexandria city	VA
Fairfax County	VA
Fairfax city	VA
Queens County	NY
Loudoun County	VA

Tabla 5. Condados-Estado con mayor MedianPrice

En la Figura 7 se observan los 10 condados con el mayor valor de MedianPrice, mientras que en la Tabla 5 se puede ver a qué estados pertenecen estos condados. A partir de esta información, se puede concluir que los condados con los precios de vivienda más altos están ubicados en los estados de VA y NY. Si se compara esta información con la Figura 3, se puede concluir que estos condados representan los valores atípicos en cada caso, siendo en ambos casos valores mayores a los 400,000 USD. Esto refuerza la observación de que New York y Virginia presentan precios medianos elevados, contribuyendo significativamente a la variabilidad observada en los datos globales.

8.1.2 Eliminación de datos atípicos.

El conjunto de datos resultante de eliminar los datos atípicos, termina con un total de 4878 datos y 23 variables, en la Tabla 6 se presenta la nueva distribución de cantidad de datos por estado.

Estado	***Cantidad de datos***
Florida (FL)	780
Carolina del Norte (NC)	1149
New Jersey (NJ)	252
New York (NY)	647
Carolina del Sur (SC)	529
Virginia (VA)	1521

Tabla 6. Cantidad de datos por estado - luego de eliminar outliers

Al comparar los datos de la Tabla 6 con la Tabla 2, se observa que el estado de NJ conserva la misma cantidad de datos; lo cual tiene relación con los análisis realizados a partir de la Figura 3, la cual no se observan datos atípicos en ese estado. Los estados como NY y VA tuvieron una disminución de 13,04% y 4,88% , respectivamente; estos fueron los estados con mayor apreciación de datos atípicos y variabilidad, de acuerdo con la Figura 3 y la Tabla 4, y a su vez son los estados con mayor impacto en la eliminación de datos atípicos.

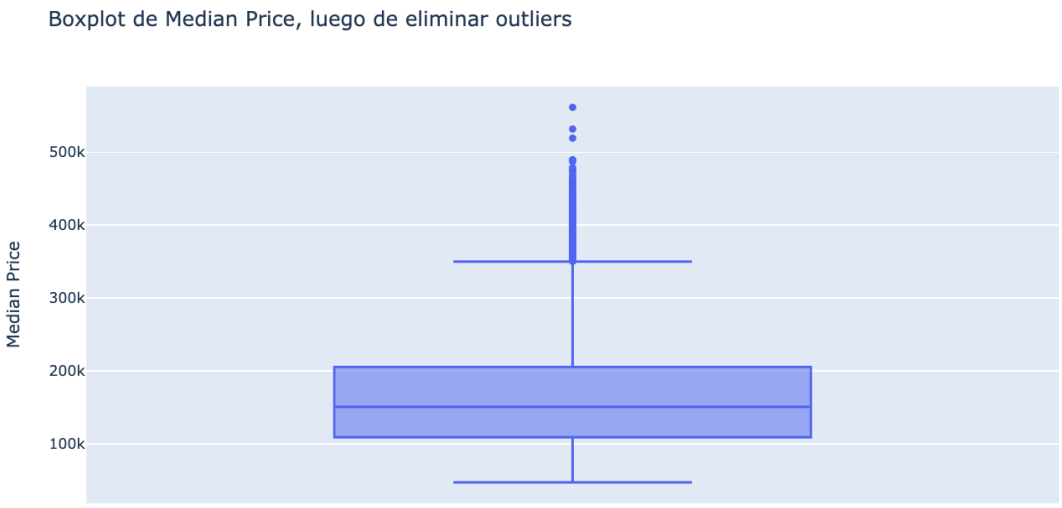

Figura 8. Boxplot Median_Price, luego de eliminar outliers

Si se compara el nuevo boxplot de la Figura 8 con el boxplot del conjunto de datos original, en la Figura 2, se notan cambios significativos en la variable *Median_Price*; mostrando una considerable distribución más ajustada, una disminución notable de datos atípicos. Los bigotes del boxplot son más cortos, lo que refleja una menor variabilidad extrema en los precios. Después de eliminar los outliers, el boxplot muestra un IQR más estrecho, indicando una reducción en la dispersión de los datos. La mediana se mantiene alrededor de 200.000 USD; sin embargo, está más centrada dentro del rango intercuartílico, lo que puede indicar una distribución más simétrica de los datos. Los bigotes del boxplot van desde, aproximadamente, los 50.000 USD hasta 500.000 USD; adicionalmente se nota, una reducción en la cantidad de outliers y mejoría en la simetría en la distribución de los datos.

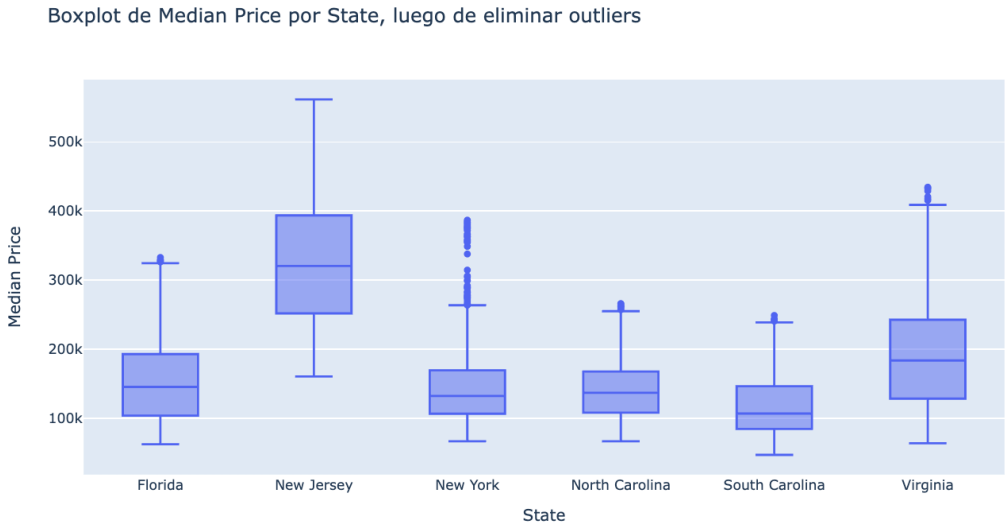

Figura 9. Boxplot Median_Price por State, luego de eliminar outliers

En la Figura 9 se observan cambios en los boxplots por estados. En comparación a la Figura 3, los estados como NY y VA siguen teniendo una dispersión mayor con respecto a los otros estados, pero con una reducción significativa de outliers. Los estados VA y NJ mantienen los precios medianos más altos, y los precios medianos más bajos siguen estando en los estados FL, NC y SC.

Para los estados de SC, NC, y FL se muestra una mejoría en la simetría después de eliminar los outliers. Las líneas medianas se observan más centradas dentro de las cajas intercuartil, indicando que los datos están más equilibrados alrededor de la mediana, y los bigotes se extienden de manera más uniforme, lo que puede sugerir una disminución en la dispersión extrema. En el caso del estado de NY, a pesar de reducir considerablemente la cantidad de outliers, se sigue observano cierta asimetría en los datos.

State	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
Florida	$154.192,44	$60.525,68	$62.400	$103.975	$145.400	$192.700	$332.800
New Jersey	$320.015,48	$90.402,96	$160.500	$252.175	$320.450	$393.575	$561.500
New York	$151.030,60	$65.688,37	$66.800	$106.650	$132.300	$169.450	$386.800
North Carolina	$142.210,18	$44.294,21	$66.800	$108.200	$136.900	$167.700	$265.900
South Carolina	$116.423,06	$41.318,70	$47.100	$84.600	$106.900	$146.300	$248.900
Virginia	$193.313,02	$78.385,40	$63.900	$128.400	$183.600	$242.600	$434.500

Tabla 7. Distribución Median Price por Estado, luego de eliminar outliers

De acuerdo con la Tabla 7, con la eliminación de los datos atípicos el valor máximo pasa a corresponder del estado de NY al estado de NJ; el cual sigue siendo el estado con mayor el precio mediano, precio minimo y el mayor precio promedio, siendo ahora seguido por el estado de VA , como el estado con el segundo valor más grande en el valor mediano. Se destaca que los estados de NY y VA disminuyen considerablemente su desviación estándar; sin embargo, siguen siendo los estados con mayor variabilidad de la variable objetivo.

8.1.3 Análisis de Componentes Principales (PCA)

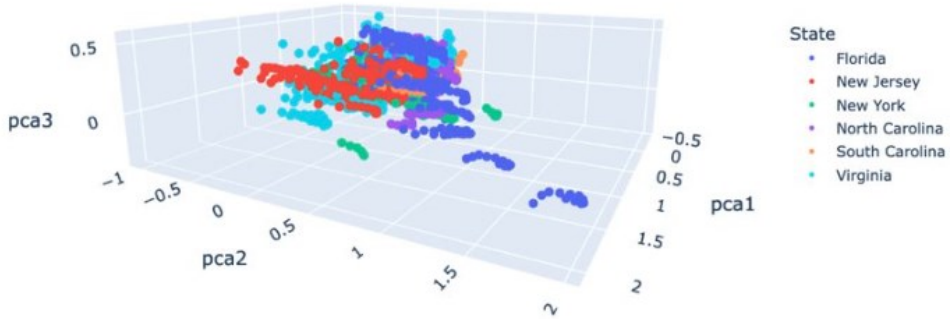

Figura 10. PCA

En la Figura 10 se muestra el agrupamiento de componentes principales, donde se destaca que los estados de Florida y Nueva York exhiben una dispersión mayor, en los componentes principales, que el resto de los estados; esto sugiere que las variables asociadas a estos estados tienen un comportamiento significativamente diferente. Por otro lado, estados como Carolina del Norte, Virginia, Carolina del Sur y New Jersey parecen tener comportamientos más similares, dado que los agrupamientos correspondientes están más cercanos entre sí. Además, esta proximidad entre algunos agrupamientos podría indicar cierto nivel de correlación entre las variables originales. En este análisis se incluyó la variable *Year,* ya que ayuda a capturar la variabilidad temporal, mejorando la separación y agrupamiento de los estados en función de sus características a lo largo del tiempo.

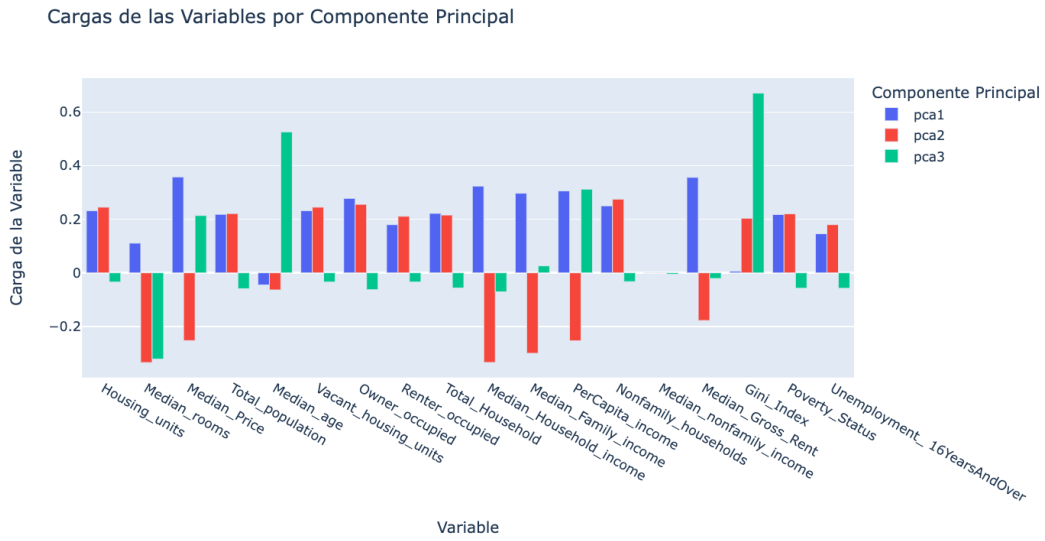

Figura 11. Cargas de las Variables por Componente Principal

De la Figura 11 se puede concluir que el primer componente principal (pca1) está fuertemente influenciado por variables como *Median_Household_Income, Median_Family_income, PerCapita_Income, Median_Price y Median_Gross_Rent;* sin embargo, presenta carga negativas en la variable *Median_age.* Lo anterior puede sugerir que esta componente tiene información de áreas con altos ingresos de la población, altos precios medianos en las viviendas y con una población con menor edad promedio.

Por otro lado, el segundo componente principal (pca2) destaca por las cargas elevadas en *Housing_units, Total_population, Vacant_housing_units y Nonfamily_households*. Sin embargo, presenta cargas negativas en las variables *Median_rooms, Median_Household_Income y Median_nonfamily_households*. Lo que puede sugerir que este componente tiene información de áreas con una mayor cantidad de casas, hogares no familiares y una población total más alta; que a su vez tienden a tener menores ingresos y viviendas con menor cantidad promediod de habitaciones.

El tercer componente principal (pca3) muestra cargas significativas en *Median_age, PerCapita_income, y Gini_Index*. Presenta cargas negativas en variables como *Median_rooms, Total_population y Owner_occupied*. Lo que sugiere, que en esta dimensión, las áreas con una mayor ingreso per cápita, mayor edad promedio tienen da tener menor cantidad de población, de viviendas ocupadas por sus propietarios y menor cantidad de promedio de habitaciones por vivienda.

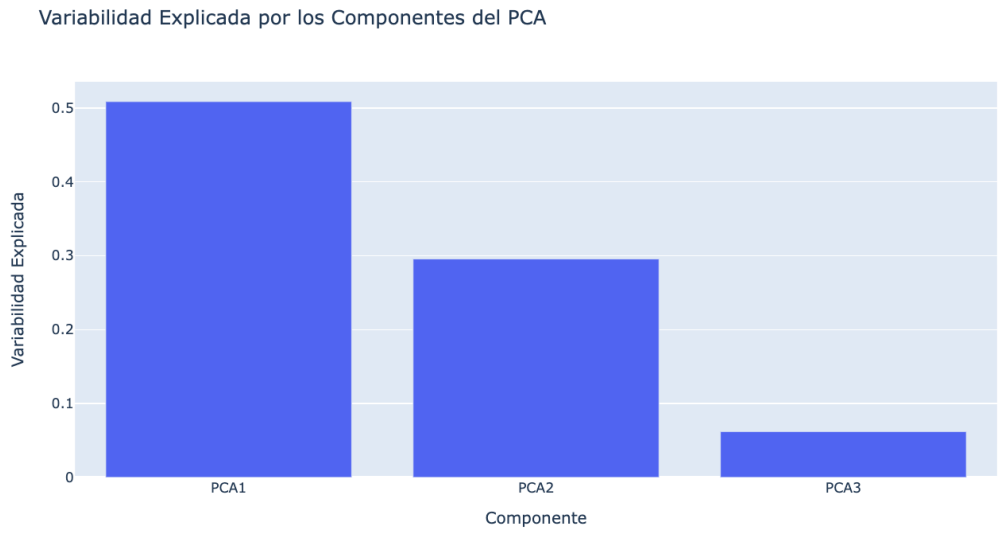

Figura 12. Variabilidad Explicada por los Componentes del PCA

La Figura 12 muestra la variabilidad explicada por cada uno de los componentes principales del PCA. El primer componente principal (pca1) explica aproximadamente el 51% de la variabilidad total, el segundo componente principal (pca2) explica alrededor del 29%, y el tercer componente principal (pca3) explica aproximadamente el 6%. Estos tres componentes juntos explican cerca del 86% de la variabilidad total en los datos, lo cual es un indicativo fuerte de que los patrones y relaciones más importantes en el dataset están bien representadoss por estos tres componentes.

8.1.4 Análisis de correlaciones

El mapa de correlación de Pearson de la Figura 13 muestra correlaciones positivas entre la mayoría de las variables; sin embargo, se ven correlaciones negativas entre Gini_Index con Median_Rooms y Median_Household_income. Adicionalmente, la variable objetivo *Median_Price* muestra correlaciones positivas con la mayoría de las variables independientes a excepción de *Median_age y Gine_Index.*

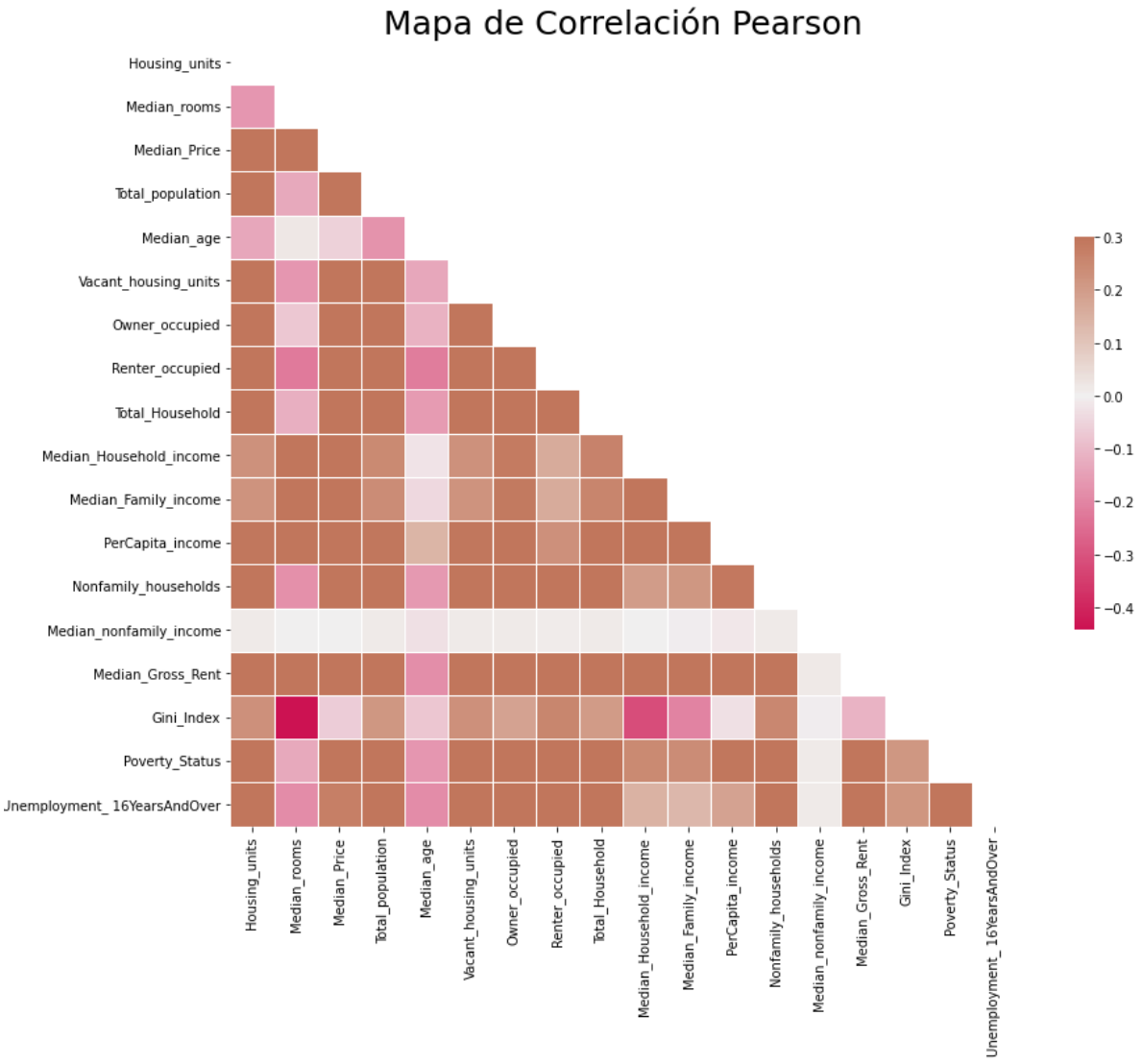

Figura 13. Mapa correlación Pearson

Para analizar las correlaciones específicas de la variable objetivo *Median_Price*, se calculan las correlaciones Pearson y Spearman de las posibles variables explicativas y la variable objetivo como se muestra en las figuras Figura 14 y Figura 15, respectivamente.

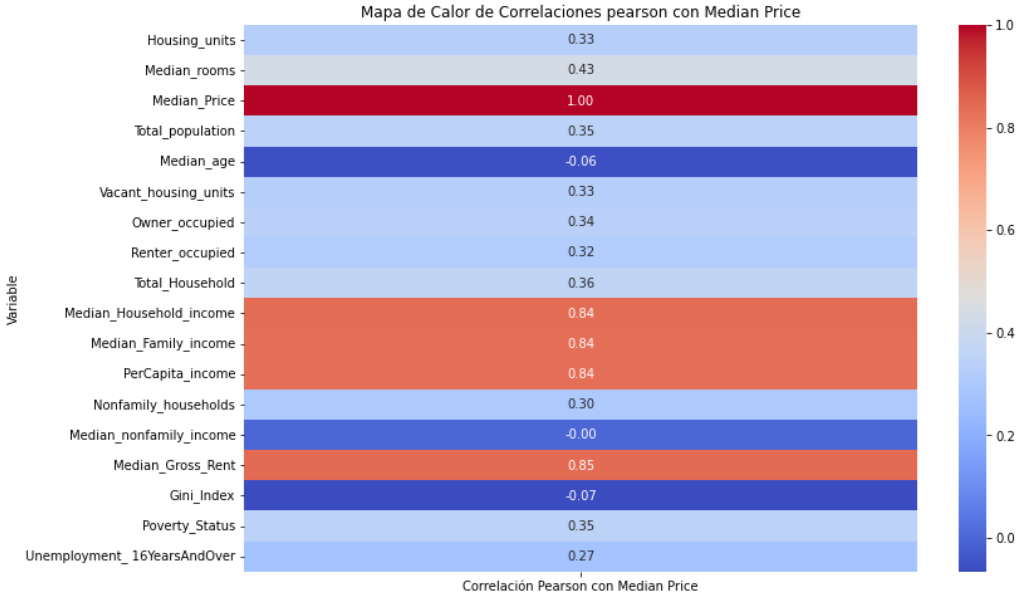

Figura 14. Correlación Pearson MedianPrice

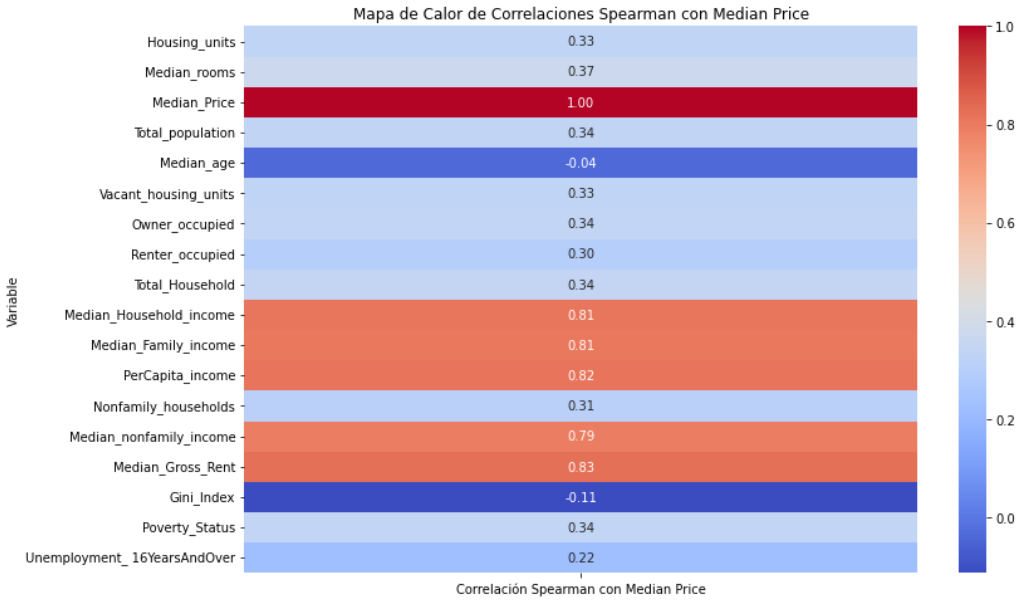

Figura 15. Correlación Spearman MedianPrice

De las figuras anteriores se concluye que hay una correlación notable entre las variables *Median_Household_income, Median_Family_income, Median_Gross_Rent y PerCapita_income* con respecto a la variable objetivo *Median_Price*. Dado que son altamente significativas en ambos métodos, podemos concluir que estas variables mantienen una relación lineal monotónica positiva. Esto nos sugiere, junto con la definición de estas variables en la Tabla 1, donde estas se relacionan con datos de ingresos por hogares o familias, que la economía familiar o de los hogares podría influir en el precio promedio de las viviendas en un área

específica. Por otro lado, las variables que se refieren a características de la ocupación de las viviendas, como *Total_population, Poverty_Status, Housing_units, Vacant_housing_units, Owner_occupied, Renter_occupied y Total_Household*, muestran correlaciones positivas en ambos métodos, pero no son lo suficientemente significativas como para establecer una relación lineal o monotónica.

8.1.5 Análisis de dependencias lineales y multicolinealidad

Para evaluar la dependencia lineal entre variables se ajusta un modelo de regresión lineal para la variable objetivo *Median_Price*, considerando cada una de las variables independientes por separado. Se evalúa cómo se ajusta cada variable mediante el coeficiente de determinación R^2. Adicionalmente, se calcula el índice de correlación multiple (VIF), para cada variable independiente, con el objetivo de evaluar la multicolinealidad entre estas. Los resultados para cada variable ajustada se presentan en la Tabla 8.

Variable	*R2*	*VIF*
Housing_units	0,8326	5,9723
Median_rooms	0,8320	5,9525
Total_population	0,8307	5,9064
Median_age	0,8325	5,9713
Vacant_housing_units	0,8326	5,9723
Owner_occupied	0,8326	5,9723
Renter_occupied	0,8326	5,9723
Total_Household	0,8326	5,9723
Median_Household_income	0,8326	5,9720
Median_Family_income	0,8292	5,8558
PerCapita_income	0,8292	5,8533
Nonfamily_households	0,8326	5,9723
Median_nonfamily_income	0,8326	5,9723
Median_Gross_Rent	0,7995	4,9884
Gini_Index	0,8320	5,9512
Poverty_Status	0,8315	5,9343
Unemployment_ 16YearsAndOver	0,8300	5,8817

Tabla 8. Valores R2 y VIF para dependencias lineales - Variables Originales

De la tabla anterior se observa que casi todas las variables tienen un coeficiente R^2 mayor a 0.8 , lo que sugiere que las variables independientes podrían estar altamente correlacionadas entre sí, duplicando información en el modelo y sugiriendo un posible sobreajuste. Además, el tener valores de VIF mayores a 5, lo que indica que las variables pueden ser expresadas como una combinación lineal

de otras. Esta situación, además de generar problemas de sobreajuste, reduciría la precisión en las predicciones de un modelo, lo que afectaría la confiabilidad en los resultados obtenidos.

Debido a la similitud en la correlación de las variables *Median_Household_income, Median_Gross_Rent, Median_Family_income, Median_Nonfamily_income* y *PerCapita_income* con la variable objetivo *Median_Price*, como se observa en la Figura 14Figura 15 , y considerando la similitud en las definiciones de cada una según la Tabla 1, se lleva a cabo un análisis de correlación entre las variables *Median_Household_income, Median_Family_income, Median_Nonfamily_income* respecto a *PerCapita_income*; esto se debe a que esta última muestra una relación más sólida con la variable objetivo *Median_Price*.

Variable	***Correlacion Pearson con PerCapita_income***
Median_Household_income	0,90
Median_Family_income	0,93
Median_Nonfamily_income	-0,01
Median_Gross_Rent	0,78
PerCapita_income	1,00

Tabla 9. Correlacion Pearson Variables Income con PerCapita_income

De acuerdo con los datos de la Tabla 9, las variables *Median_Household_income, Median_Gross_Rent y Median_Family_income* tienen una correlación significativamente alta con respecto a *PerCapita_income*, lo cual significa que que estas variables pueden ser explicadas por la variable *PerCapita_income*.

Adicionalmente, considerando la similitud en las definiciones de cada una según la Tabla 1, se lleva a cabo un análisis de correlación entre las variables *Vacant_housing_units, Owner_occupied y Renter_occupied* respecto a *Housing_units;* ya que esta ultima variable, por definición y relación con la variable objetivo, engloba más información.

Variable	***Correlacion Pearson con Housing_units***
Vacant_housing_units	1,00
Owner_occupied	0,97
Renter_occupied	0,93
Housing_units	1,00

Tabla 10. Correlacion Pearson Variables housing con Housing_units

De la Tabla 10 se puede conlquir que las variables *Vacant_housing_units, Owner_occupied y Renter_occupied* tienen una correlación significativamente alta con respecto a *Housing_units*, lo cual significa que que estas variables pueden ser explicadas por la variable *Housing_units*.

8.2 SELECCIÓN DE CARACTERÍSTICAS

8.2.1 Usando correlación tradicional

En esta técnica se calcula la correlación Pearson entre las variables predictoras y la variable objetivo. Se seleccionan aquellas variables cuya correlación sea inferior a 0.8 y las resultantes conforman un nuevo conjunto de datos a evaluar.

Con esta técnica se eliminan 4 variables del conjunto original, dando como resultado una mejoría en los coeficientes y valores VIF; lo que sugiere una mayor independencia entre las variables seleccionadas. Adicionalmente, al entrenar y evaluar una regresión lineal múltiple, se obtiene un $RMSE$ de 0,111 y un R^2 de 0,4493; y se obtienen como variables no significativas *Median_age, Poverty_Status y Median_nonfamily_income.* El detalle de estos resultados se encuentra en la sección **¡Error! No se encuentra el origen de la referencia.**.

Al usar esta técnica de selección de variables y establecer el umbral de 0.8, se estarían eliminando las variables: *Median_Gross_Rent, Median_Family_income, Median_Household_income y PerCapita_income.* Estas variables, cuyas definiciones están alineadas todas con un contexto económico de la población como se ve en la Tabla 1, han demostrado ser importantes en la predicción de precios de vivienda, como se menciona en el *PLANTEAMIENTO DEL PROBLEMA* y en el *ESTADO DEL ARTE*. Si se eliminan todas estas, se estaría perdiendo información crucial para la predicción, lo cual se ve reflejado en el R^2 obtenido que muestra una moderada explicación en la variabilidad de la variable objetivo. Por ello, es esencial evaluar otros métodos de selección de características para asegurarse de que no se sacrifica la precisión del modelo.

8.2.2 Usando índice de correlación múltiple

En este caso se utiliza el Factor de Inflación de la Varianza (VIF), para evaluar la multicolinealidad entre las variables predictoras. Se calcula el VIF para cada variable predictora y se establece un umbral de 5, de manera que aquellas con valor VIF a 5 no quedaran en el conjunto de variables seleccionadas ya que un VIF superior a 5 indica que una variable está altamente correlacionada con otras variables del modelo, lo que puede inflar los errores de los coeficientes estimados y disminuir la precisión del modelo.

Al usar esta técnica se conservaron 4 de las 17 variables, obteniendo valores de R^2 y VIF significativamente menores en comparación con el conjunto de datos original y el conjunto seleccionado mediante correlación tradicional. Se obtiene un $RMSE$ de 0,1313 y un R^2 de 0,2390, al entrenar y evaluar una regresión lineal con este conjunto de variables.

Al usar esta técnica de selección de características, se eliminaron más variables en comparación con la selección por correlación tradicional. Esto sugiere que el índice de correlación múltiple es más estricto en la eliminación de variables redundantes, lo cual puede ser beneficioso para construir un modelo más simple. Sin embargo, también es importante considerar que eliminar demasiadas variables puede resultar en la pérdida de información crucial para la predicción, lo cual se observa eb el bajo % de explicación de variabilidad obtenido.

8.2.3 Usando forward

Esta técnica consta de un proceso iterativo que inicia con un modelo vacío y en cada iteración se evalúa la significancia estadística de las variables restantes, añadiendo la variable con el valor p más bajo al modelo. Este proceso se repite hasta que no se puedan añadir más variables que mejoren significativamente el modelo. Este enfoque incluye solo aquellas variables que aportan información significativa para la predicción, lo que puede ayudar a reducir el riesgo de multicolinealidad y mejora la interpretabilidad de un posible modelo de predicción de la variable objetivo.

En el uso de esta técnica se eliminan 8 variables. Se conserva una mayor cantidad de variables en comparación con las técnicas anteriores; adicional de tener una mejor precisión al entrenar y evaluar regresión lineal, con un $RMSE$ de 0,00625 y un R^2 de 0,8277. Sin embargo, es importante destacar la presencia de alta multicolinealidad, que puede ser explicada al incluirse variables como *Median_Gross_Rent, Median_Family_income y PerCapita_income*, las cuales todas tienen una alta correlación con la variable objetivo como se observa en la Figura 14; adicionalmente, se debe tener en cuenta que las variables *Median_Gross_Ren y, Median_Family_income* están altamente correlacionadas con *PerCapita_income.*

8.2.4 Usando backward

La selección de características usando backward consiste también en un proceso iterativo; sin embargo, a diferencia de la técnica forward, en este caso se comienza con un modelo que incluye todas las variables predictoras y en cada iteración se elimina la variable estadísticamente menos significativa. Este proceso se repite hasta que todas las variables restantes en el modelo sean estadísticamente significativas. La técnica forward puede incluir variables que inicialmente parecen significativas, pero luego resultan no serlo cuando se agregan más variables, mientras que backward considera el impacto de todas las variables desde el principio.

El conjunto de variables seleccionadas por esta técnica conserva 11 variables del conjunto original, se obtiene un $RMSE$ de 0,0614 y un R^2de 0,9334. Dentro de las variables eliminadas están *Housing_units y Median_nonfamily_income; e*stos

resultados son similares a los de la técnica forward. Adicionalmente en ambos casos se conservan variables altamente correlacionadas, esto puede influir en la multicolinealidad y confiabilidad de un posible modelo de predicción usando estos conjuntos de variables

8.2.5 Usando Lasso Regression

La regresión Lasso (Least Absolute Shrinkage and Selección Operator) se utiliza comúnmente en conjuntos de datos que presentan alta dimensionalidad y problemas de sobreajuste [37].Este modelo hace uso de la regularización L1, en la que se añade un término a la función de perdida al modelo para penalizar la magnitud de los coeficientes de las variables haciendo uso del encogimiento o shrinkage. Este método hace que el modelo Lasso sea efectivo para realizar la predicción y la selección de características, ya que elimina las características irrelevantes [36]. Para usar la regresión Lasso como técnica de selección de características o variables, se entrena el modelo y luego se hace extraen las variables que fueron conservadas por el modelo al ser consideradas importantes para la predicción de la variable objetivo.

Esta técnica tiene resultados comparables con forward y backward; sin embargo, elimina más variables que la técnica backward y sigue presentando un alta multicolinealidad, lo cual se debe a que conserva las variables *Median_Gross_Rent, Median_Family_income, Median_Household_income y PerCapita_income*.

8.2.6 Usando Random Forest

Para la selección de características usando Random Forest, se entrena un RandomForestRegressor con parámetros n_estimators=100 y random_state=42. Una vez entrenado el modelo, se utiliza la clase *SelectFromModel* para identificar y seleccionar las características más importantes para la predicción de la variable objetivo, según su importancia en el modelo de Random Forest.

Se eliminan 13 de las 17 variables originales, esto puede simplificar el modelo y reducir la multicolinealidad. Sin embargo, se puede perder información importante que puede ayudar a mejorar la precisión del modelo y afecta la capacidad de generalización. En comparación con los otros métodos que conservan más variables, y presentan un R^2 más alto en la regresión lineal multiple con los conjuntos de variables seleccionados, el modelo resultante de la selección por Random Forest puede no capturar completamente todas las complejidades y relaciones que pueden existir en los datos originales, afectando la precisión y confiabilidad en la predicción de los precios de las viviendas.

8.2.7 Conjunto de variables seleccionadas

La selección de características debe considerar aspectos como la mejora de la multicolinealidad del conjunto de variables original, la simplicidad de un posible modelo y la necesidad de tener información suficiente para capturar la variabilidad y complejidad de los datos. Con base en esto y en los resultados encontrados en el análisis y comparación de las técnicas, se decide tomar el conjunto de variables seleccionadas usando la técnica backward debido a sus resultados y a que esta técnica considera el impacto de todas las variables desde el principio, eliminando aquellas que no contribuyen significativamente a la predicción. Sin embargo, se decide eliminar, adicionalmente, las variables *Median_Gross_Rent y Median_Family_income* que tienen una alta correlación con la variable objetivo y a su vez con la variable *PerCapita_income*, por lo que todas pueden influir de manera similar en la predicción de la variable objetivo.

Se decide mantener *PerCapita_income*, ya que de acuerdo a su definición puede abarcar información de manera más general y fue seleccionada por la mayoría de las técnicas de selección de variables, demostrando así su importancia e influencia en la predicción de los precios de las viviendas. A este nuevo conjunto de variables seleccionadas se le aplica la misma evaluación que a los conjuntos resultantes por las técnicas anteriormente expuestas.

Variable	*R2*	*VIF*
Poverty_Status	0,7610	4,1841
Total_population	0,7610	4,1840
Total_Household	0,7608	4,1804
Median_rooms	0,7607	4,1791
Unemployment_ 16YearsAndOver	0,7594	4,1567
Gini_Index	0,7583	4,1376
Renter_occupied	0,7496	3,9939
Nonfamily_households	0,7310	3,7174
PerCapita_income	0,4077	1,6882

Tabla 11. Valores R2 y VIF - Variables seleccionadas

Se puede observar en la Tabla 11 que en este nuevo conjunto de variables seleccionadas, el R^2 está entre 0,73 y 0,76, los valores de VIF están entre 3,7 y 4,2 para la mayoría de las variables, indicando una mejor independencia entre las variables en comparación con los altos valores de VIF observados en el conjunto original y en la selección por backward, donde muchas variables presentaban valores de VIF superiores a 5. En el caso especifico de la variable *PerCapita_income*, se presenta una disminución en el R^2 y en el valor VIF bajando a 0,41 y 1,69, respectivamente; indicando que, al eliminar las otras variables

altamente correlacionadas se ha reducido su redundancia y mejorado su valor explicativo.

Variable	*Coefficients*	*P-values*
const	0,0285	1,55E-03
Median_rooms	-0,0473	1,40E-03
Total_population	2,3774	2,57E-02
Renter_occupied	0,7373	1,27E-34
Total_Household	0,3037	2,37E-01
PerCapita_income	0,9849	0,00E+00
Nonfamily_households	-1,1386	3,80E-96
Gini_Index	-0,0851	5,63E-12
Poverty_Status	-2,0503	5,26E-02
Unemployment_ 16YearsAndOver	0,3015	8,27E-08

Tabla 12. Coeficientes y p-values regresión lineal - Variables seleccionadas

Al entrenar y evaluar una regresión lineal multiple con este nuevo conjunto de datos, se obtienen un RMSE de 0,0742 y un R^2 de 0,76. Adicionalmente, de acuerdo con los resultados de la Tabla 12, las variables que se consideran estadísticamente no significativas son *Total_Household y Poverty_Status.*

A pesar de que este nuevo conjunto de variables presenta un R^2 menor en la regresión lineal multiple, en comparación con el obtenido usando el conjunto de variables seleccionadas por backward; al haber eliminado variables altamente correlacionadas, se mejora la independencia de las variables restantes, dando lugar a un conjunto menos redudante y un modelo más confiable y con mayor capacidad de generalización. Por tanto, este nuevo conjunto de datos puede ser utilizado eficazmente para la predicción del precio mediano de las viviendas.

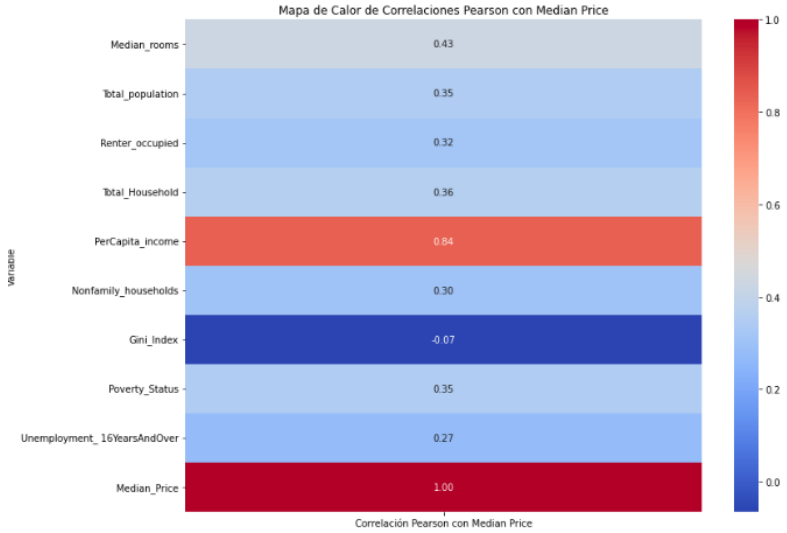

Figura 16. Correlación Pearson MedianPrice - Conjunto final

En la Figura 16 se observa que la mayoría de las variables tiene una correlación que no es significativamente alta con respecto a la variable objetivo Median_Price, a excepción de PerCapita_Income; sin embargo, como se ha expuesto anteriormente, la eliminación de esta variable implica pérdida de información importante lo que se ve reflejado en un R^2 menor, como es el caso del conjunto de variables seleccionadas en el punto 8.2.1. En la fase de modelación se usa la evaluación de las métricas $RMSE$ y R^2 de los conjuntos de entrenamiento, prueba y validación, con el objetivo de evaluar que el modelo no presente sobreajuste y confirmar la generalización del modelo ante datos no vistos previamente.

8.3 MODELACIÓN Y OPTIMIZACIÓN DE HIPERPARÁMETROS

8.3.1 Regresión lineal multiple

De acuerdo con los resultados y conclusiones obtenidas en el conjunto de variables seleccionadas, se deciden eliminar las variables *Total_Household* y *Poverty_Status* que mostraron ser estadísticamente no significativas al ser incluidas en las variables para entrenar y evaluar una regresión lineal multiple, de acuerdo con los p-values de la Tabla 12.

Model	*rmse_train*	*r2_train*	*rmse_test*	*r2_test*	*rmse_val*	*r2_val*
LinearRegression	0,0590	0,8350	0,0636	0,8314	0,0714	0,8294

Tabla 13. Métricas evaluación - Regresión lineal multiple

Se puede observar en la Tabla 13 que la regresión lineal cuenta con bajos errores y altos R^2, indicando que puede explicar aproximadamente el 83% de la variabilidad tanto en el conjunto de entrenamiento, testeto y validación. La baja diferencia entre

las métricas de los diferentes conjuntos evaluados, muestra que generaliza bien a nuevos datos y minimiza el sobreajuste.

8.3.2 Modelos con parámetros por defecto

Model	*rmse_train*	*r2_train*	*rmse_test*	*r2_test*	*rmse_val*	*r2_val*
RandomForestRegressor	0,014	0,991	0,068	0,808	0,099	0,673
Lasso	0,145	0,000	0,158	-0,035	0,191	-0,222
KNeighborsRegressor	0,034	0,945	0,060	0,849	0,086	0,753
GradientBoostingRegressor	0,041	0,919	0,063	0,833	0,083	0,770
AdaBoostRegressor	0,068	0,779	0,092	0,647	0,108	0,610
XGBRegressor	0,007	0,997	0,055	0,874	0,086	0,752
KerasRegressor (ANN)	0,039	0,929	0,056	0,870	0,073	0,822

Tabla 14. Métricas evaluación - Modelos con parámetros por defecto

En la Tabla 14 se observa que el modelo Lasso presenta $RMSE$ alto y R^2 bajo o negativo en los tres conjuntos de datos, lo que indica que el modelo no está capturando la variabilidad y no es adecuado para este conjunto de datos. Los modelos XGBRegressor (XGB), GradientBoostingRegressor (GB) y KNeighborsRegressor (KNN), tienen bajos errores y altos R^2 en el conjunto de entrenamiento; sin embargo, estos valores se diferencian significativamente a las métricas obtenidas en los conjuntos de prueba y validación, indicando sobreajuste en estos modelos y menor capacidad de generalización a datos nuevos. Estas diferencias en las métricas es más grande en los modelos RandomForestRegressor (RF) y AdaBoostRegressor (AB), concluyendo que en este caso estos modelos presentan mayor sobreajuste.

En el caso del modelo ANN se obtienen menores diferencias entre las métricas de los conjuntos de entrenamiento, prueba y validación, en comparación a los otros modelos; por lo que se puede concluir que el modelo generaliza bien a datos no vistos previamente y es capaz de capturar relaciones complejas, teniendo un bajo error y alta explicación de la variabilidad.

8.3.3 Modelos y optimización de hiperparámetros usando RandomizedSearchCV

Model	*best_params*	*rmse_ train*	*r2_train*	*rmse_ test*	*r2_test*	*rmse_ val*	*r2_val*
RandomForest Regressor	{'criterion': 'mae', 'max_depth': 45, 'max_features': 5, 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 79}	0,013	0,992	0,057	0,865	0,083	0,767
Lasso	{'alpha': 45}	0,145	0,000	0,158	-0,035	0,191	-0,222
KNeighbors Regressor	{'n_neighbors': 2, 'p': 1, 'weights': 'distance'}	0,000	1,000	0,052	0,889	0,096	0,689
GradientBoosting Regressor	{'learning_rate': 0.1, 'max_depth': 8, 'n_estimators': 197, 'subsample': 0.5}	0,004	0,999	0,057	0,866	0,088	0,741
AdaBoost Regressor	{'learning_rate': 1.0, 'n_estimators': 57}	0,068	0,783	0,091	0,655	0,107	0,617
XGB Regressor	{'colsample_bytree': 0.5, 'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 150, 'subsample': 1.0}	0,013	0,992	0,053	0,881	0,079	0,790
Keras Regressor (ANN)	{'activation': 'relu', 'batch_size': 17, 'epochs': 25, 'loss': 'mae', 'neurons': 29, 'optimizer': 'adam', 'patience': 3}	0,047	0,894	0,059	0,853	0,076	0,809

Tabla 15.Métricas evaluación - Modelos con hiperparámetros optimizados usando RandomizedSearchCV

En la Tabla 15 se observa que la optimización de hiperparámetros mediante RandomizedSearchCV mejora el desempeño de algunos modelos comparado con los resultados obtenidos sin optimizar. En el caso de los modelos RF, KNN y XGB mejoran sus métricas en el conjunto de validación, en el caso de RF y KNN también mejoran en el conjunto de prueba; indicando que la optimización encuentra un conjunto de parámetros que ayuda a estos modelos a aumentar su capacidad de explicación de la variable objetivo. Sin embargo, sigue habiendo sobreajuste al tener una significante diferencia entre el desempeño de los conjuntos de entrenamiento, prueba y validación. Por otro lado, los modelos GB y ANN muestran una leve disminución en su desempeño en los conjuntos de prueba y validación; en ambos casos sigue existiendo sobre ajuste, menor capacidad de generalización y predicción a datos más recientes. El modelo Lasso y el modelo AB no muestran una

diferencia significativa en comparación al desempeño obtenido sin optimizar los hiperparámetros.

8.3.4 Modelos y optimización de hiperparámetros usando GridSearchCV

Model	*best_params*	*rmse_train*	*r2_train*	*rmse_test*	*r2_test*	*rmse_val*	*r2_val*
RandomForest Regressor	{'criterion': 'mae', 'max_depth': 20, 'max_features': 3, 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 50}	0,012	0,993	0,054	0,881	0,081	0,779
Lasso	{'alpha': 0.05}	0,093	0,586	0,094	0,629	0,105	0,632
KNeighbors Regressor	{'n_neighbors': 3, 'p': 1, 'weights': 'distance'}	0,000	1,000	0,052	0,886	0,093	0,709
GradientBoosting Regressor	{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.5}	0,015	0,989	0,052	0,887	0,078	0,796
AdaBoost Regressor	{'learning_rate': 0.2, 'n_estimators': 200}	0,069	0,778	0,093	0,640	0,110	0,594
XGB Regressor	{'colsample_bytree': 0.5, 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200, 'subsample': 0.5}	0,008	0,997	0,051	0,893	0,074	0,815
Keras Regressor (ANN)	{'activation': 'relu', 'batch_size': 16, 'epochs': 50, 'loss': 'mse', 'neurons': 32, 'optimizer': 'rmsprop', 'patience': 5}	0,046	0,890	0,057	0,862	0,061	0,821

Tabla 16. Métricas evaluación - Modelos con hiperparámetros optimizados usando GridSearchCV

En la Tabla 16 se observa que la optimización de hiperparámetros mejora el desempeño de algunos modelos comparado con los resultados obtenidos sin optimizar y los resultados obtenidos optimizando los hiperparámetros usando RandomizedSearchCV, para otros modelos se mantiene o se empeora su desempeño. El modelo Lasso mejora considerablemente su desempeño en todos los conjuntos, mostrando que esta optimización da con un valor alpha que le permite explicar el 63% de la variabilidad de los precios medianos de las viviendas. Los modelos RF, KNN, GB y XGB mejoran sus métricas en el conjunto de validación y en el conjunto de prueba; indicando que esta optimización encuentra un conjunto de parámetros que ayuda a estos modelos a aumentar su capacidad de explicación de la variable objetivo y su capacidad de generalización ante datos no vistos. Sin embargo, sigue habiendo sobreajuste al tener diferencias entre el desempeño de los conjuntos de prueba y validación con respecto al conjunto de entrenamiento. El modelo ANN también mejora sus métricas en conjuntos de datos no vistos previamente, disminuyendo así las diferencias con respecto al conjunto de entrenamiento y mejorando el sobreajuste presentado en los resultados obtenidos sin optimziar hiperparámetros y los resultados obtenidos optimizando los hiperparámetros usando RandomizedSearchCV.

8.4 ELECCIÓN DEL MEJOR MODELO

Modelos como XBG y GB a pesar de tener un alto porcentaje de ajuste y bajo error, tienen considerables diferencias en las métricas de los diferentes conjuntos de datos; indicando que pueden no generalizar tan bien como el modelo ANN o la regresión lineal, cuyas diferencias entre las métricas es menor.

Con base en las métricas obtenidas en el entrenamiento, prueba y validación de diferentes modelos, se define que el mejor modelo es la regresión lineal múltiple. Este modelo cuenta con bajos $RMSE$ y altos R^2 en los conjuntos de entrenamiento, prueba y validación, indicando que tiene un alto porcentaje de explicación de la variable objetivo, una alta precisión y tiene las menores diferencias entre las métricas de los conjuntos. Se puede concluir que presenta el menor sobreajuste y mayor capacidad de generalización a datos no vistos previamente.

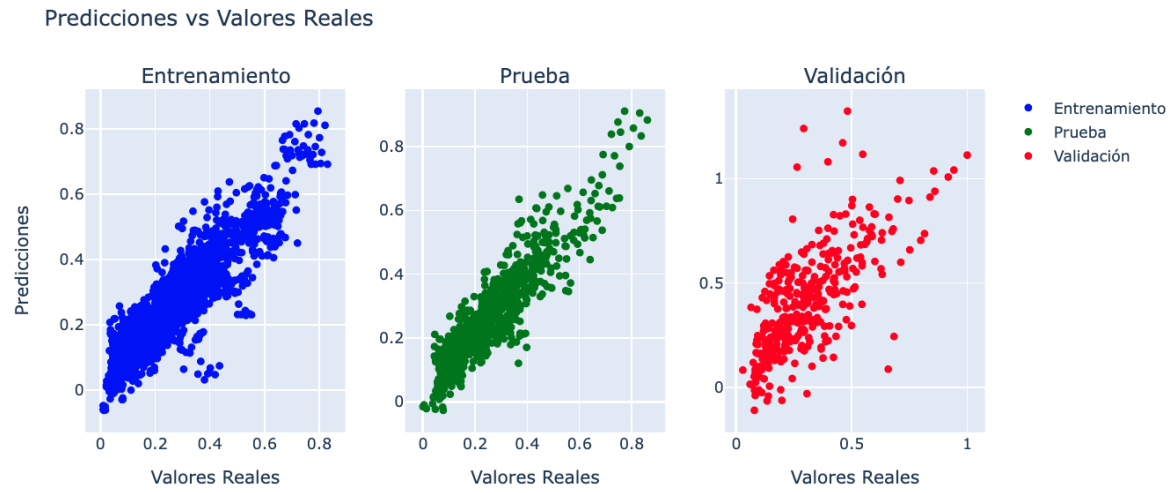

Figura 17. Valores reales vs Valores de la predicción de Median_Price

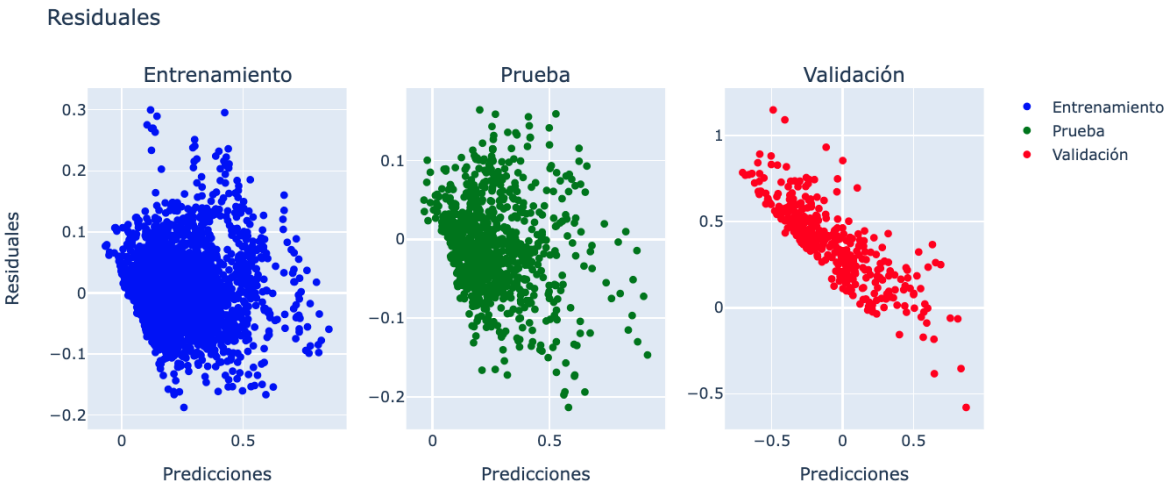

Figura 18. Residuales en la predicción de Median_Price

En la Figura 17 se observan las predicciones vs los valores reales de Median_Price. Para el conjunto de entrenamiento, las predicciones se alinean de manera muy cercana a los valores reales, lo cual concuerda con el valor $RMSE$ bajo de 0,05 y un alto R^2 de 0,84 en el entrenamiento del modelo. Para el conjunto de prueba, las predicciones también se alinean bien con los valores reales, aunque se nota una ligera dispersión mayor en comparación con el conjunto de entrenamiento, este conjunto obtuvo un $RMSE$ de 0,063 y un R^2 de 0,83. En el conjunto de validación las predicciones siguen alineándose adecuadamente con los valores reales, concordando con el $RMSE$ de 0,071 y el R^2 de 0,83; aunque se nota mayor dispersión que en los conjuntos de entrenamiento y validación. Sin embargo, se puede confirmar que el modelo generaliza bien a datos no vistos previamente.

Se observa en la Figura 18 que en el conjunto de entrenamiento los residuales están centrados alrededor de cero y no muestran un patrón claro. Para el conjunto de prueba los residuales también están centrados alrededor de cero, aunque hay una mayor variabilidad en comparación con el conjunto de entrenamiento; lo que indica que el modelo es confiable y sigue teniendo un buen desempeño en datos no vistos previamente. En el conjunto de validación los residuales muestran una mayor dispersión, con residuales que muestran una tendencia descendente conforme aumentan las predicciones. Aunque el modelo generaliza bien a datos no vistos previamente, esta tendencia indica que podría haber factores adicionales que no están siendo capturados adecuadamente por el modelo.

Variable	***Coefficients***	***P-values***
PerCapita_income	1,101	0,00E+00
Renter_occupied	0,859	5,94E-86
Total_population	0,615	1,86E-22
State_New Jersey	0,056	2,40E-22
State_Virginia	0,042	4,73E-30
State_North Carolina	0,007	2,99E-02
Year_encoded	-0,009	5,24E-90
State_South Carolina	-0,011	8,53E-03
Gini_Index	-0,039	2,14E-04
State_New York	-0,050	1,48E-35
Unemployment_ 16YearsAndOver	-0,078	1,03E-01
Median_rooms	-0,101	3,41E-13
Nonfamily_households	-1,040	2,41E-120

Tabla 17. Coeficientes y p-values regresión - modelo final

A partir de la Tabla 17 se concluye que la variable *Unemployment_ 16YearsAndOver* se considera estadísticamente no significativa en la predicción de los precios medianos de viviendas. A continuación se muestra la importancia de las variables para la predicción de la variable objetivo.

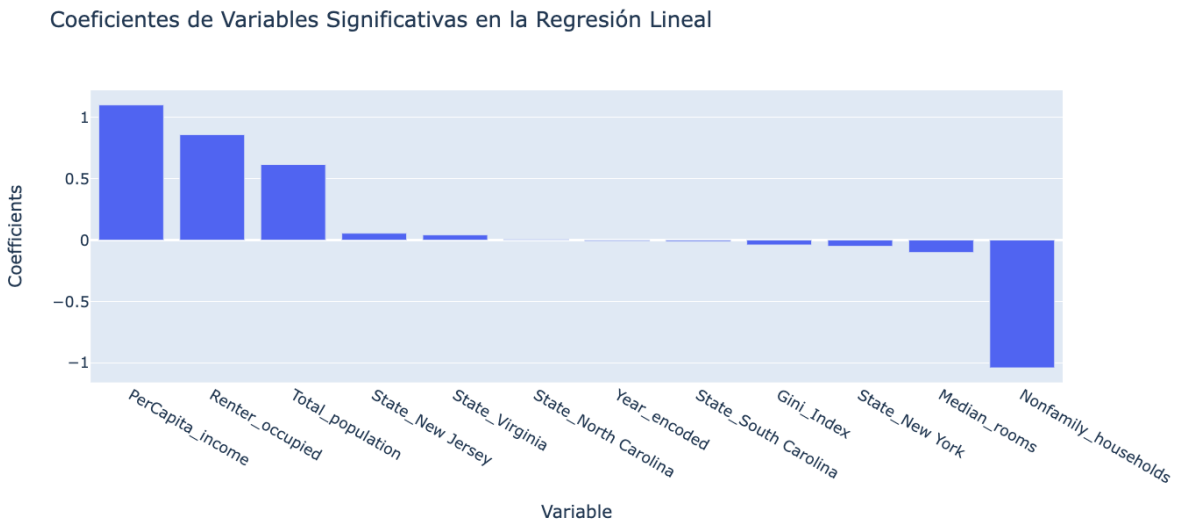

Figura 19. Importancia de las variables usadas en la predicción de Median_Price

A partir de la Figura 19 se puede concluir que la variable con más influencia en la predicción de la variable objetivo es PerCapita_income, al tener el coeficiente positivo más alto indica que un aumento en su valor está fuertemente asociado con un aumento en el precio mediano de las viviendas. Las variables *Renter_occupied y Total_population* también presentan coeficientes positivos significativos, sugiriendo que estos factores contribuyen positivamente al precio mediano de las viviendas. Por otro lado, la variable *Nonfamily_households* tiene el coeficiente negativo más alto, indicando que un aumento en la proporción de hogares no familiares está asociado con una disminución en el precio mediano de mediano; al igual que la cantidad promedio de habitaciones por vivienda , ya que la variable *Median_rooms* tiene un coeficiente negativo. Las variables dummies relacionadas a los estados son significativas a pesar de tener coeficientes más pequeños, lo que sugiere que las diferencias entre los estados capturadas por las variables dummies tienen un impacto en la predicción del precio de las viviendas.

Otras variables, como *Gini_Index y Year* , presentan coeficientes más pequeños, indicando un impacto menor en la predicción del precio mediano. A pesar de ser estadísticamente significativas, tienen un efecto menos pronunciado en comparación con las variables mencionadas anteriormente.

9 CONCLUSIONES

Resumen de resultados:

A partir de los análisis y resultados expuestos en este documento, se puede concluir que para la predicción de los precios medianos de las viviendas en zonas costeras en EE.UU., es importante tener en cuenta las características de los estados en los que se encuentran ubicadas las viviendas, adicional de los ingresos económicos de la población, el total de la población y la cantidad de viviendas ocupadas por inquilinos.

Para poder obtener las variables con mayor impacto y aporte en la predicción de los precios medianos de las viviendas, es importante aplicar y evaluar varias técnicas de selección de características al conjunto de datos original, de manera que el conjunto seleccionado explique la variable objetivo y no presente multicolinealidad. Sin embargo, también es crucial combinar esta información con decisiones basadas en el contexto y conocimiento del problema que se quiere solucionar, ya que esto permite evaluar si existe alguna variable relevante y comúnmente usada que se esté quedando fuera del conjunto de variables a usar. Esto se observó al seleccionar las características mediante el uso de la técnica backward, y adicionalmente al eliminar variables altamente correlacionadas pero manteniendo la información de PerCapita_income. Con este enfoque se obtuvo un conjunto de datos que mejoró la multicolinealidad y a su vez contenía información suficiente para explicar la variabilidad de los precios inmobiliarios en ciudades costeras del Atlántico en EE. UU.

En predicción de los precios medianos de las viviendas en zonas costeras en EE.UU, los modelos de ML como Random Forest, Gradient Boosting y XGBoost, demostraron potencial en la predicción de precios inmobiliarios en ciudades costeras del Atlántico en EE. UU. Se beneficiaron de la optimización de hiperparámetros, mejorando su desempeño en términos de mejores métricas y menor diferencia entre estas en los conjuntos de datos evaluados. Sin embargo, presentaron problemas de sobreajuste, indicando menor capacidad de generalización y explicación de la variabilidad en datos previamente no vistos.

El uso de técnicas de deep learning, como redes neuronales, en la predicción de los precios medianos de las viviendas en zonas costeras en EE.UU mostró potencial, obteniendo mejores métricas y menor sobreajuste que los modelos de ML; indicando que generaliza bien a datos no vistos previamente y es capaz de capturar relaciones complejas, teniendo un bajo error y alta explicación de la variabilidad. Sin embargo, es un modelo con mayor gasto computacional que la regresión lineal múltiple, la cual también obtuvo un buen desempeño en las métricas de $RMSE$ y R^2; adicionalmente, muestra una menor diferencia entre las métricas de los conjuntos de entrenamiento, prueba y validación, indicando una mayor capacidad de generalización y menor sobreajuste, siendo un modelo más simple y con mayor capacidad de interpretabilidad.

Finalmente, se puede concluir que para predecir los precios promedio de bienes inmuebles en estados con zonas costeras del Atlántico en Estados Unidos, es fundamental tener un conjunto de datos proveniente de diferentes contextos como el demográfico y socioeconómico; junto con una adecuada de selección de variables y técnicas de modelado, ajustadas y evaluadas en el contexto específico del problema.

Implicaciones prácticas:
Los modelos de predicción del precio promedio de vivendas, presentados en este documento, pueden servir como herramienta para estimar el valor de las viviendas en diferentes escenarios en estados con zonas costeras del Atlántico en Estados Unidos. Los resultados obtenidos pueden ser utilizados por diversos agentes interesados, siendo información beneficiosa para tomar decisiones informadas sobre compra y venta de viviendas en los estados analizados.

Limitaciones del estudio:
La disponibilidad de los datos históricos es una limitación significativa en la predicción de precios en el mercado inmobiliario; presentando inconvenientes no solo por la poca cantidad de registros, sino también en la diversidad de variables que se pueden obtener en las fuentes de datos actualmente. Esto es crucial, ya que la precisión de los modelos se ve directamente afectada por la cantidad y calidad de los datos usados en el entrenamiento de estos, influenciando en los resultados finales.

Recomendaciones para futuras investigaciones:
Es recomendable ampliar el conjunto de datos usados, en términos de obtener más registros e incluir más variables; no solo del contexto socioeconómico y demográfico, sino también otros contextos como clima, políticas y otros. que permitan capturan información que pueda explicar el comportamiento del mercado inmobiliario en zonas costeras. Adicionalmente se puede ahondar más en la integración de técnicas avanzadas como el Deep Learning, ya que el uso de este tipo de modelos mostró un buen desempeño que podría ser mejorado para aumentar la precisión en la predicción de los precios de las viviendas en ciudades costeras. Este tipo de técnicas puede obtener mejores resultados siendo usadas en conjuntos con mayor cantidad de datos y variables.

Impacto del uso del Machine Learning en sector inmobiliario: En este trabajo se confirma el potencial de las técnicas de Machine Learning para la predicción de precios del sector inmobiliario en zonas costeras. Contribuyendo en herramientas para estimar el valor de las viviendas, y con esto contribuir a que sea un mercado con un flujo equitativo de información para todos.

REFERENCIAS

[1] N. K. Kishor, “Forecasting House Prices: The Role of Fundamentals, Credit Conditions, and Supply Indicators,” The Journal of Real Estate Finance and Economics, 2023, doi: 10.1007/s11146-023-09971-y.

[2] S. Jia, Y. Wang, and G.-Z. Fan, “Home-Purchase Limits and Housing Prices: Evidence from China,” The Journal of Real Estate Finance and Economics, vol. 56, no. 3, pp. 386–409, 2018, doi: 10.1007/s11146-017-9615-2.

[3] Z. Sun and J. Zhang, “Research on Prediction of Housing Prices Based on GA-PSO-BP Neural Network Model: Evidence from Chongqing, China,” International Journal of Foundations of Computer Science, 2022, doi: 10.1142/S0129054122420163.

[4] J.-S. Chou, D.-B. Fleshman, and D.-N. Truong, “Comparison of machine learning models to provide preliminary forecasts of real estate prices,” Journal of Housing and the Built Environment, 2022, doi: 10.1007/s10901-022-09937-1.

[5] M. Koetter and T. Poghosyan, “Real estate prices and bank stability,” J Bank Financ, vol. 34, no. 6, pp. 1129–1138, 2010, doi: 10.1016/j.jbankfin.2009.11.010.

[6] D. Tchuente and S. Nyawa, “Real estate price estimation in French cities using geocoding and machine learning,” Ann Oper Res, vol. 308, no. 1–2, pp. 571–608, 2022, doi: 10.1007/s10479-021-03932-5.

[7] A. M. B. Pedersen, A. Weissensteiner, and R. Poulsen, “Financial planning for young households,” Ann Oper Res, vol. 205, no. 1, pp. 55–76, 2013, doi: 10.1007/s10479-012-1205-3.

[8] Y. Kang et al., “Understanding house price appreciation using multi-source big geo-data and machine learning,” Land use policy, vol. 111, 2021, doi: 10.1016/j.landusepol.2020.104919.

[9] M. Talaga, M. Piwowarczyk, M. Kutrzyński, T. Lasota, Z. Telec, and B. Trawiński, Apartment valuation models for a big city using selected spatial attributes, vol. 11683 LNAI. 2019. doi: 10.1007/978-3-030-28377-3_30.

[10] J. Vonlanthen, “Interest rates and real estate prices: a panel study,” Swiss J Econ Stat, vol. 159, no. 1, 2023, doi: 10.1186/s41937-023-00111-0.

[11] D. Epley, “A better method to estimate price change in single family housing: A test of median-to-median compared to repeat sales,” International Journal of Housing Markets and Analysis, vol. 5, no. 4, pp. 377–391, 2012, doi: 10.1108/17538271211268529.

[12] J. D. Fisher and S. R. Rutledge, “The impact of Hurricanes on the value of commercial real estate,” Business Economics, vol. 56, no. 3, pp. 129–145, 2021, doi: 10.1057/s11369-021-00212-9.

[13] L. Jaitman, “Urban infrastructure in Latin America and the Caribbean: Public policy priorities Research at the policy frontier in Latin America: Health, Education, Infrastructure and Housing and Climate Change Sebastian

Galiani," Lat Am Econ Rev, vol. 24, no. 1, 2015, doi: 10.1007/s40503-015-0027-5.
[14] J. Beck and M. Lin, "Impacts of sea level rise on real estate prices in coastal Georgia," Review of Regional Studies, vol. 50, no. 1, pp. 43–52, 2020.
[15] A. Soltani, M. Heydari, F. Aghaei, and C. J. Pettit, "Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms," Cities, vol. 131, 2022, doi: 10.1016/j.cities.2022.103941.
[16] K. Cao, M. Diao, and B. Wu, "A Big Data–Based Geographically Weighted Regression Model for Public Housing Prices: A Case Study in Singapore," Ann Am Assoc Geogr, vol. 109, no. 1, pp. 173–186, 2019, doi: 10.1080/24694452.2018.1470925.
[17] M. Bussas, C. Sawade, N. Kühn, T. Scheffer, and N. Landwehr, "Varying-coefficient models for geospatial transfer learning," Mach Learn, vol. 106, no. 9–10, pp. 1419–1440, 2017, doi: 10.1007/s10994-017-5639-3.
[18] E. C. M. Hui, C. K. Chau, L. Pun, and M. Y. Law, "Measuring the neighboring and environmental effects on residential property value: Using spatial weighting matrix," Build Environ, vol. 42, no. 6, pp. 2333–2343, 2007, doi: 10.1016/j.buildenv.2006.05.004.
[19] M. A. Salles, P. O. Canziani, and R. H. Compagnucci, "Spatial Variations in the Average Rainfall-Altitude Relationship in Great Britain: An Approach using Geographically Weighted Regression," International Journal of Climatology, vol. 21, no. 4, pp. 455–466, 2001, doi: 10.1002/joc.614.
[20] J. C. Viriato, "AI and machine learning in real estate investment," Journal of Portfolio Management, vol. 45, no. 7, pp. 43–54, 2019, doi: 10.3905/jpm.2019.45.7.043.
[21] A. Louati, R. Lahyani, A. Aldaej, A. Aldumaykhi, and S. Otai, "Price forecasting for real estate using machine learning: A case study on Riyadh city," Concurr Comput, vol. 34, no. 6, 2022, doi: 10.1002/cpe.6748.
[22] L. Yang, Y. Liang, Q. Zhu, and X. Chu, "Machine learning for inference: using gradient boosting decision tree to assess non-linear effects of bus rapid transit on house prices," Ann GIS, vol. 27, no. 3, pp. 273–284, 2021, doi: 10.1080/19475683.2021.1906746.
[23] Y. Zhao, G. Chetty, and D. Tran, "Deep Learning with XGBoost for Real Estate Appraisal," in 2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019, 2019, pp. 1396–1401. doi: 10.1109/SSCI44817.2019.9002790.
[24] G. Srirutchataboon, S. Prasertthum, E. Chuangsuwanich, P. N. Pratanwanich, and C. Ratanamahatana, "Stacking Ensemble Learning for Housing Price Prediction: A Case Study in Thailand," in KST 2021 - 2021 13th International Conference Knowledge and Smart Technology, 2021, pp. 73–77. doi: 10.1109/KST51265.2021.9415771.
[25] T. Lasota, E. Sawiłow, B. Trawi Ń Ski, M. Roman, P. Marczuk, and P. Popowicz, A method for merging similar zones to improve intelligent models for real estate appraisal, vol. 9011. 2015. doi: 10.1007/978-3-319-15702-3_46.

[26] A. Farahani, B. Pourshojae, K. Rasheed, and H. R. Arabnia, “A Concise Review of Transfer Learning,” Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.02144
[27] T.-T. Nguyen and S. Yoon, “A novel approach to short-term stock price movement prediction using transfer learning,” Applied Sciences (Switzerland), vol. 9, no. 22, 2019, doi: 10.3390/app9224745.
[28] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” IEEE Trans Inf Theory, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.
[29] L. Breiman, “Random forests,” Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[30] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” Front Neurorobot, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
[31] Y. Freund and R. E. Schapire, “A Short Introduction to Boosting,” 1999. [Online]. Available: www.research.att.com/fyoav,
[32] J. Fan et al., “Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China,” Energy Convers Manag, vol. 164, pp. 102–111, 2018, doi: 10.1016/j.enconman.2018.02.087.
[33] H. Li, Z. Zhang, and Z. Liu, “Application of artificial neural networks for catalysis: A review,” Catalysts, vol. 7, no. 10, 2017, doi: 10.3390/catal7100306.
[34] H. Akoglu, “User’s guide to correlation coefficients,” Turk J Emerg Med, vol. 18, no. 3, pp. 91–93, Sep. 2018, doi: 10.1016/J.TJEM.2018.08.001.
[35] H. Liu, “Chapter 7 - Description methods of spatial wind along railways,” in Wind Forecasting in Railway Engineering, H. Liu, Ed., Elsevier, 2021, pp. 251–282. doi: https://doi.org/10.1016/B978-0-12-823706-9.00007-7.
[36] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: Theory and practice,” Neurocomputing, vol. 415, pp. 295–316, 2020, doi: https://doi.org/10.1016/j.neucom.2020.07.061.
[37] M. R. Putri, I. G. P. S. Wijaya, F. P. A. Praja, A. Hadi, and F. Hamami, “The Comparison Study of Regression Models (Multiple Linear Regression, Ridge, Lasso, Random Forest, and Polynomial Regression) for House Price Prediction in West Nusa Tenggara,” in ICADEIS 2023 - International Conference on Advancement in Data Science, E-Learning and Information Systems: Data, Intelligent Systems, and the Applications for Human Life, Proceeding, 2023. doi: 10.1109/ICADEIS58666.2023.10270916.
[38] S. Naik and A. Puthran, “Metro Cities House Price Prediction Using ML,” in International Conference on Self Sustainable Artificial Intelligence Systems, ICSSAS 2023 - Proceedings, 2023, pp. 588–594. doi: 10.1109/ICSSAS57918.2023.10331796.

ANEXOS

1. Selección de características

1.1 Usando correlación tradicional

Variable	*R2*	*VIF*
Vacant_housing_units	0,4185	1,7197
Nonfamily_households	0,4185	1,7197
Housing_units	0,4185	1,7197
Owner_occupied	0,4185	1,7197
Renter_occupied	0,4185	1,7197
Total_Household	0,4185	1,7197
Poverty_Status	0,4185	1,7197
Median_nonfamily_income	0,4184	1,7195
Median_age	0,4179	1,7180
Total_population	0,4172	1,7158
Gini_Index	0,4108	1,6972
Unemployment_ 16YearsAndOver	0,4083	1,6899
Median_rooms	0,2314	1,3010

Tabla 18. Valores R2 y VIF - selección usando corr

Al comparar los resultados proporcionados por la Tabla 18 con los resultados del conjunto de datos original en la Tabla 8 , se observa que las variables seleccionadas usando el umbral de 0,8 en la correlación tradicional, tienen coeficientes R^2 y valores de VIF significativamente menores en comparación con el conjunto de datos original. Por ejemplo, variables como *Total_Household, Owner_occupied, Housing_units, y Nonfamily_households* tienen un R^2 de aproximadamente 0,41 y un VIF de alrededor de 1,7.

Una regresión lineal multiple entrenada y evaluada con este conjunto de variables seleccionadas, explica aproximadamente el 44,9% de la variabilidad en los precios de las viviendas, mostrando una capacidad moderada de predicción. A continuación, se obtienen los coeficientes y p-values de cada variable usada en este conjunto de datos con el objetivo de identificar cuáles de ellas no son estadísticamente significativas para el modelo.

Variable	*Coefficients*	*P-values*
const	-0,1193	3,13E-01
Housing_units	0,9981	1,36E-12
Median_rooms	0,6606	4,92E-208
Total_population	-3,8134	3,52E-02

Median_age	-0,0218	2,62E-01
Vacant_housing_units	0,9981	1,36E-12
Owner_occupied	0,7033	2,47E-07
Renter_occupied	2,4490	4,67E-49
Total_Household	3,1848	2,71E-39
Nonfamily_households	-2,0665	4,56E-46
Median_nonfamily_income	-0,0307	7,93E-01
Gini_Index	0,1329	2,85E-12
Poverty_Status	-0,9266	6,08E-01
Unemployment_ 16YearsAndOver	-0,7163	5,57E-16

Tabla 19. Coeficients y p-values regresión lineal - variables seleccionadas usando corr

De la Tabla 19 se concluye que las variables *Median_age, Poverty_Status y Median_nonfamily_income* no son significativamente estadísticas al tener un p-vapue mayor a 0.05. Por otro lado, variables como *Median_rooms, Renter_occupied y Nonfamily_households* son altamente significativas y contribuyen de manera relevante a la predicción del precio de las viviendas.

1.2 Usando índice de correlación múltiple

Variable	*R2*	*VIF*
Median_nonfamily_income	0,2031	1,2548
Median_age	0,2001	1,2501
Gini_Index	0,1854	1,2277
Median_rooms	0,0081	1,0081

Tabla 20. Valores R2 y VIF - selección usando indice de correlación múltiple

En la Tabla 20 se observan los resultados de la regresión lineal múltiple utilizando el conjunto de variables seleccionadas por el índice de correlación múltiple. Variables como *Median_nonfamily_income y Median_age,* que el conjunto original mostraban un R^2 cercano a 0,83 y VIF de alrededor de 5,9, con la selección por el índice de correlación multiple presentan un R^2 de aproximadamente 0,2 y un VIF de alrededor de 1,25, lo que indica una menor multicolinealidad y una mayor independencia entre las variables seleccionadas.

Variable	*Coefficients*	*P-values*
const	-0,1026	4,57E-01
Median_rooms	0,5961	3,47E-166
Median_age	-0,0603	2,39E-03
Median_nonfamily_income	0,0065	9,62E-01
Gini_Index	0,1988	4,14E-20

Tabla 21. Coeficientes y p-values regresión lineal - variables seleccionadas usando vif

Se observa en la Tabla 21 que la variable *Median_nonfamily_income* tiene un p-value mayor a 0.05, lo que sugiere que no es estadísticamente significativa en el modelo y su inclusión podría estar introduciendo ruido, reduciendo la precisión del modelo. En este caso, variables como *Median_rooms y Gini_Index* tienen p-values menores a 0.05, lo que indica que son altamente significativas y contribuyen de manera relevante a la predicción del precio de las viviendas.

Las variables eliminadas con esta tenica de selección de características son: *Vacant_housing_units, Nonfamily_households, PerCapita_income, Median_Gross_Rent, Total_Household, Total_population, Poverty_Status, Renter_occupied, Median_Family_income, Housing_units, Unemployment_ 16YearsAndOver, Median_Household_income, Owner_occupied.*

1.3 Usando forward

Variable	*R2*	*VIF*
Median_age	0,8291	5,8499
Gini_Index	0,8288	5,8397
Median_rooms	0,8284	5,8288
Unemployment_ 16YearsAndOver	0,8261	5,7492
Median_Family_income	0,8250	5,7155
PerCapita_income	0,8233	5,6592
Renter_occupied	0,8156	5,4245
Nonfamily_households	0,7997	4,9926
Median_Gross_Rent	0,7844	4,6392

Tabla 22. Valores R2 y VIF - selección usando forward

Las variables seleccionadas presentan R^2 altos, cercanos a 0,82, y valores de VIF alrededor de 5; mostrando una alta multicolinealidad en comparación con los conjuntos de variables seleccionadas con la correlación tradicional y usando el índice de correlación multiple. Sin embargo, el hecho de que más variables hayan sido seleccionadas sugiere que puede haber una influencia significativa de estas en la predicción del precio de las viviendas, que debe ser confirmada con la evaluación de la significancia estadística.

Al entrenar y evaluar una regresión lineal múltiple utilizando el conjunto de variables seleccionadas por esta tecnica, se obtiene un $RMSE$ de 0,0625 y un R^2 de 0,8277. Lo que indica que el modelo explica aproximadamente el 82,7% de la variabilidad en la variable objetivo; siendo mejor en comparación con los resultados de las técnicas anteriores, lo cual sugiere que este conjunto de variables seleccionadas tiene mejor capacidad predictiva entre los métodos evaluados.

Variable	*Coefficients*	*P-values*
const	-0,0539	1,23E-06
PerCapita_income	0,3714	3,44E-30
Median_Gross_Rent	0,4608	7,63E-188
Median_Family_income	0,3026	9,53E-21
Renter_occupied	0,7020	3,93E-58
Nonfamily_households	-0,7767	3,61E-127
Unemployment_ 16YearsAndOver	0,3781	1,59E-18
Median_rooms	-0,0630	2,18E-06
Gini_Index	0,0402	6,81E-04
Median_age	0,0251	4,80E-02

Tabla 23. Coeficientes y p-values regresión lineal - variables seleccionadas usando forward

En la Tabla 23 se observa que todas las variables seleccionadas mediante forward presentan significancia estadística, indicando una fuerte relación con la variable objetivo.

1.4 Usando backward

Variable	*R2*	*VIF*
Median_rooms	0,8315	5,9351
Gini_Index	0,8314	5,9309
Poverty_Status	0,8310	5,9167
Total_Household	0,8308	5,9119
Total_population	0,8302	5,8892
Unemployment_ 16YearsAndOver	0,8294	5,8622
Median_Family_income	0,8263	5,7571
PerCapita_income	0,8248	5,7079
Nonfamily_households	0,8192	5,5305
Renter_occupied	0,8178	5,4875
Median_Gross_Rent	0,7854	4,6596

Tabla 24. Valores R2 y VIF - selección usando backward

Con el uso de esta técnica se eliminan las variables *Median_nonfamily_income, y Housing_units ,* ambas variables fueron eliminadas también por la técnica forward. Las variables que si fueron seleccionadas, presentan valores de R^2 entre 0,78 y 0,83 y valores VIF entre 4,6 y 5,9; lo cual indica una alta multicolinealidad en este conjunto de variables, similar a lo observado en el conjunto de variables seleccionadas usando forward.

Variable	*Coefficients*	*P-values*
const	-0,0445	1,26E-08
Median_rooms	-0,0642	1,37E-06
Total_population	-5,3450	1,95E-08
Renter_occupied	0,8382	3,25E-60
Total_Household	1,0555	1,26E-06
Median_Family_income	0,3244	3,43E-28
PerCapita_income	0,3493	8,56E-40
Nonfamily_households	-0,7822	4,05E-63
Median_Gross_Rent	0,4540	3,95E-199
Gini_Index	0,0497	9,10E-06
Poverty_Status	4,1709	7,65E-06
Unemployment_ 16YearsAndOver	0,3988	4,84E-17

Tabla 25. Coeficientes y p-values regresión lineal - variables seleccionadas usando backward

En la Tabla 25 se observa que todas las variables seleccionadas mediante la técnica backward presentan significancia estadística, indicando una fuerte relación con la variable objetivo.

1.5 Usando Lasso Regression

Variable	*R2*	*VIF*
Median_Household_income	0,8293	5,8589
Median_age	0,8291	5,8500
Gini_Index	0,8289	5,8447
Median_rooms	0,8285	5,8301
Median_Family_income	0,8271	5,7839
Unemployment_ 16YearsAndOver	0,8262	5,7539
PerCapita_income	0,8243	5,6921
Renter_occupied	0,8160	5,4340
Nonfamily_households	0,8019	5,0473
Median_Gross_Rent	0,7938	4,8485

Tabla 26. Valores R2 y VIF - selección usando Lasso Regression

De la Tabla 26 se concluye que las variables seleccionadas mediante la técnica de regresión Lasso presentan R^2 cercaos a 0,82, los cuales se consideran significativamente altos, y valores de VIF que entre 4,8 y 5,8, indicando una multicolinealidad alta.

En total se eliminan 7 variables, las cuales son: *Poverty_Status, Total_Household, Total_population, Vacant_housing_units, Median_nonfamily_income,*

Housing_units y Owner_occupied. Sin embargo, a pesar de eliminar más variables que backward, se presenta igualmente una alta multicolinealidad. El conjunto de variables seleccionadas usando la regresion Lasso, al ser usado para entrenar y evaluar una regresión lineal multiple obtiene un RMSE de 0,062 y un R^2 de 0,8277.

Variable	*Coefficients*	*P-values*
const	-0,0539	1,67E-06
Median_rooms	-0,0631	5,17E-06
Median_age	0,0251	4,83E-02
Renter_occupied	0,7018	3,80E-57
Median_Household_income	0,0016	9,73E-01
Median_Family_income	0,3016	4,11E-12
PerCapita_income	0,3710	7,45E-26
Nonfamily_households	-0,7765	2,73E-119
Median_Gross_Rent	0,4605	8,92E-155
Gini_Index	0,0405	6,50E-03
Unemployment_ 16YearsAndOver	0,3779	4,72E-18

Tabla 27. Coeficientes y p-values regresión lineal - variables seleccionadas usando Lasso Regression

En la Tabla 27 se observa que la variable *Median_Household_income* tiene un p-value mayor a 0,05, por lo que se considera estadísticamente no significativa. Esto indica que no hay suficiente evidencia estadística para afirmar que esta variable tiene un impacto significativo en la predicción de la variable objetivo.

1.6 Usando Random Forest

Variable	*R2*	*VIF*
Median_Household_income	0,7966	4,9168
Median_Family_income	0,7958	4,8977
PerCapita_income	0,7824	4,5965
Median_Gross_Rent	0,7354	3,7795

Tabla 28. Valores R2 y VIF - selección usando Random Forest

Al usar la selección de variables de Random Forest se eliminan 13 de las 17 variables originales, las cuales presentan valores de R^2 entre 0,73 y 0,79, y valores de VIF de entre 3,7 y 4,9, lo cual indica una baja multicolinealidad en comparación con los métodos anteriores. Los resultados de entrenar y evaluar una regresión lineal multiple utilizando las variables seleccionadas por Random Forest, son un RMSE de 0,068 y un R^2de 0,7940. Estos valores indican una capacidad predictiva moderadamente alta del modelo, aunque inferior a los obtenidos con las técnicas de forward, backward y la selección de la regresión Lasso.

Variable	*Coefficients*	*P-values*
const	-0,0371	1,66E-37
Median_Household_income	0,0143	6,99E-01
Median_Family_income	0,1704	4,70E-05
PerCapita_income	0,3786	2,94E-57
Median_Gross_Rent	0,4891	7,22E-213

Tabla 29. Coeficientes y p-values regresión lineal - variables seleccionadas usando Random Forest

De las variables conservadas en este método, la variable *Median_Household_income* se considera estadísticamente no significativa al tener un p-value mayor a 0,05; el resto de las variables son estadísticamente significativas con p-values extremadamente bajos, lo que destaca su relevancia en la predicción del precio de las viviendas; a pesar de que han mostrado una correlación alta con la variable objetivo, en la Figura 14, y entre ellas mismas como se estableció en la Tabla 9.