|  |
| --- |
| PROGRAMACION PARA CIENCIA DE DATOS            UNIDAD 4 16 MARZO 2025  PRESENTADO A: SEBASTIAN RODRIGUEZ        FUNDACIÓN UNIVERSITARÍA COMPENSAR      PRESENTADO POR: JUAN DAVID PEPICANO C            CIENCIA DE DATOS  BOGOTÁ, D.C  2025 |
|  |

Contenidos temáticos:

Contrastes de hipótesis en la toma de decisiones.

Estimación de parámetros. Conjuntos confidenciales.

Prueba de Hipótesis

Este informe presenta un análisis estadístico sobre el precio de apartamentos en Medellín, abordando:

* Contraste de hipótesis sobre la relación entre el estrato y el precio.
* Modelos de regresión lineal y logística para predecir el precio y el estrato de los apartamentos.

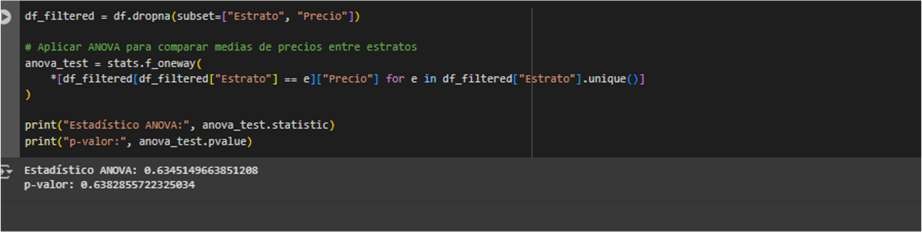
Pregunta: ¿Existe una diferencia significativa en el precio según el estrato socioeconómico?

* H₀: No hay diferencia en el precio medio entre los estratos.
* H₁: Hay una diferencia significativa en los precios entre estratos.

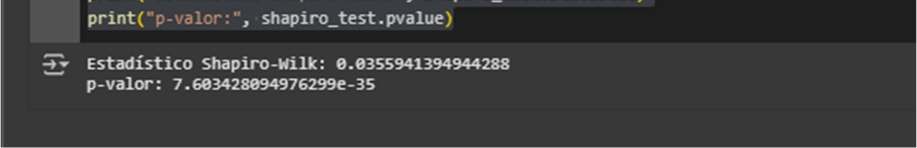
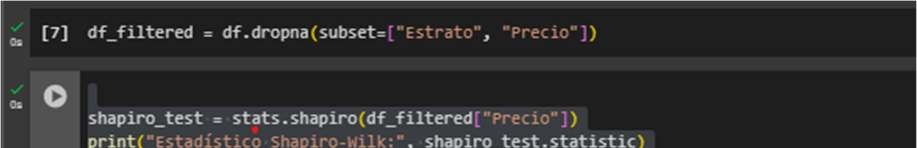
Estas hipótesis son mutuamente excluyentes (solo una puede ser verdadera) y colectivamente exhaustivas (cubren todas las posibles relaciones entre las variables).

Elección de la Prueba Estadística

Dado que queremos evaluar la relación entre una variable categórica (Estrato) y una variable continua (Precio), es apropiado utilizar una prueba t de Student para comparar las medias de dos grupos, o una ANOVA si estamos comparando más de dos grupos. Sin embargo, dado que el estrato tiene múltiples categorías, ,me decanto por por la prueba ANOVA de un factor

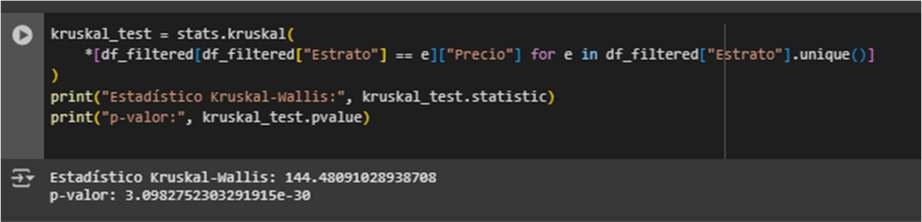


El resultado de la prueba ANOVA en la imagen muestra un p-valor de 0.638, lo que significa que no hay suficiente evidencia estadística para rechazar la hipótesis nula. En otras palabras, el precio no varía significativamente entre los diferentes estratos.



El resultado del test de Shapiro-Wilk en la imagen muestra un p-valor de 7.60e-35, lo que es mucho menor que 0.05. Esto indica que los datos no siguen una distribución normal.

usar Kruskal-Wallis, que es una alternativa no paramétrica que no requiere normalidad.



El resultado de la prueba Kruskal-Wallis en la imagen muestra un p-valor de 3.09e-30, lo que es extremadamente bajo (< 0.05).

Interpretación del resultado:

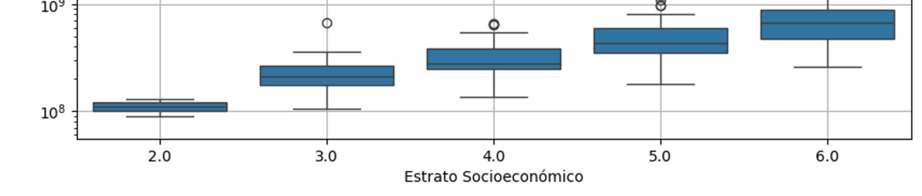
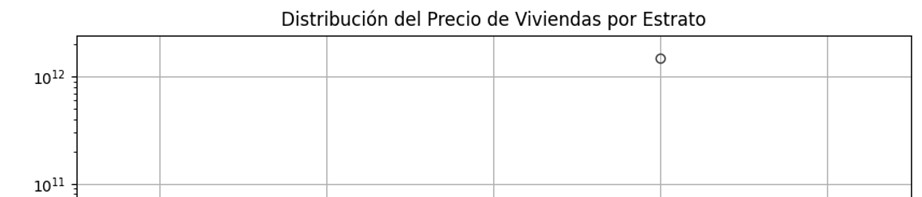
Se rechaza la hipótesis nula (H₀).

Existe una diferencia estadísticamente significativa en los precios de los apartamentos según el estrato.

El estrato influye en el precio de los apartamentos en Medellín.

Esto confirma que los precios no son aleatorios respecto al estrato y que hay una relación clara entre ambas variables.

boxplot para visualizar la distribución de los precios por estrato, con escala logarítmica para mejorar la interpretación



Conclusión del Contraste de Hipótesis

Los resultados obtenidos en la prueba de Kruskal-Wallis muestran un p-valor extremadamente bajo (< 0.05) , lo que indica que existe una diferencia significativa en los precios de los apartamentos entre los diferentes estratos socioeconómicos.

El boxplot refuerza esta conclusión, mostrando una tendencia clara en la que los precios tienden a aumentar con el estrato. Sin embargo, se observan valores atípicos en todos los estratos, lo que sugiere que existen casos excepcionales de viviendas con precios considerablemente altos dentro de cada categoría.

Dado que la hipótesis nula ha sido rechazada, podemos afirmar que el estrato tiene un impacto significativo en el precio de los apartamentos en Medellín. Este hallazgo es crucial para quienes buscan entender la dinámica del mercado inmobiliario en la ciudad y puede ser útil para inversores, compradores y reguladores del sector.

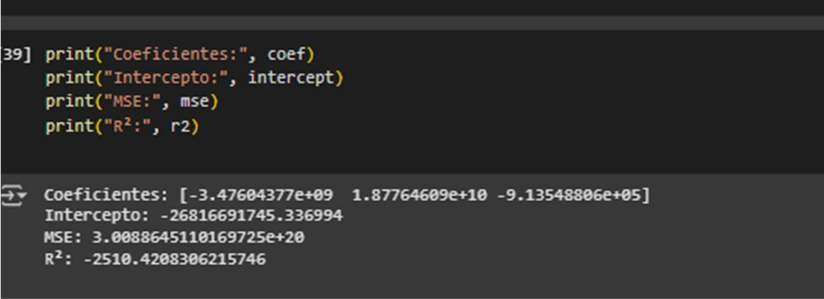
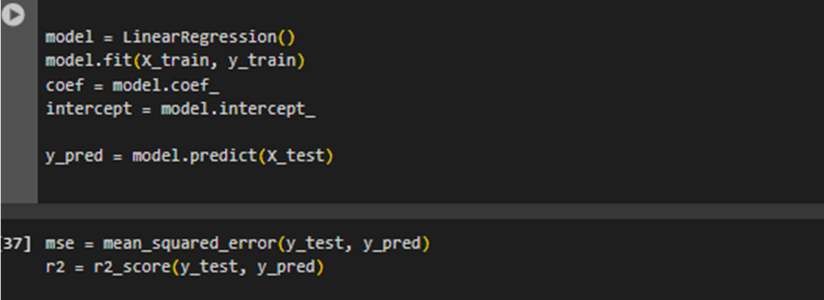
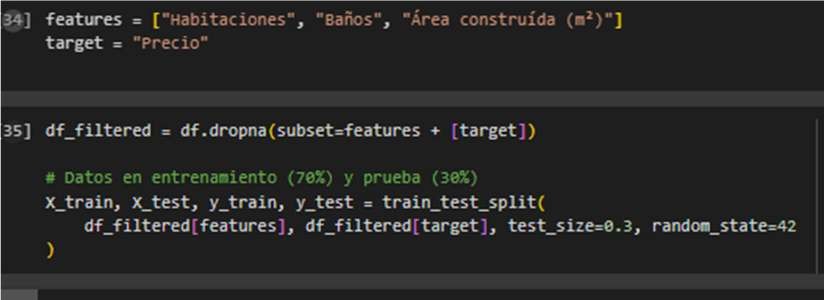
Regresión Lineal y Logística

Para complementar este análisis, utilizamos modelos de regresión lineal y logística para evaluar qué factores influyen en el precio y predecir el estrato de una vivienda en función del área construida

Regresión Lineal

Para predecir el precio de las viviendas, seleccionamos tres variables continuas independientes:

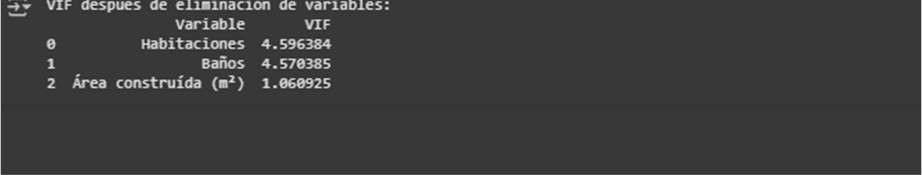
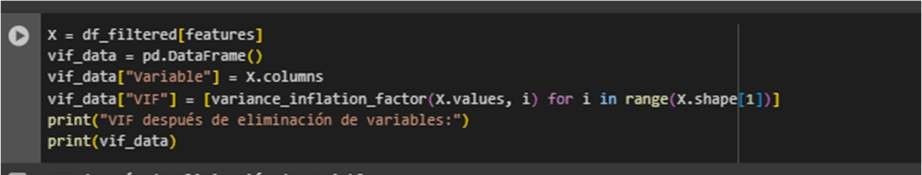
Habitaciones, Baños y Área construida (m²).



R² negativo (-2510.42), lo que indica que el modelo no solo es malo, sino que es peor que una simple media.

Dado que los resultados iniciales de la regresión lineal fueron problemáticos (R² negativo y MSE elevado), aplicamos ajustes para mejorar el modelo:

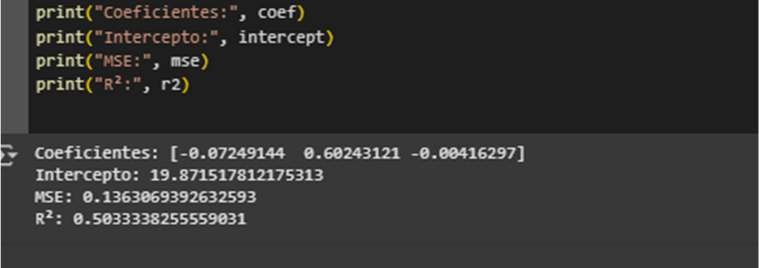
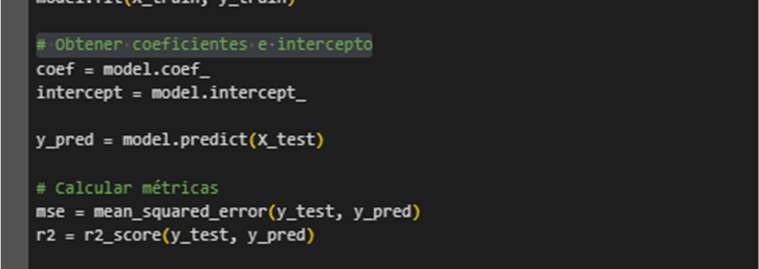
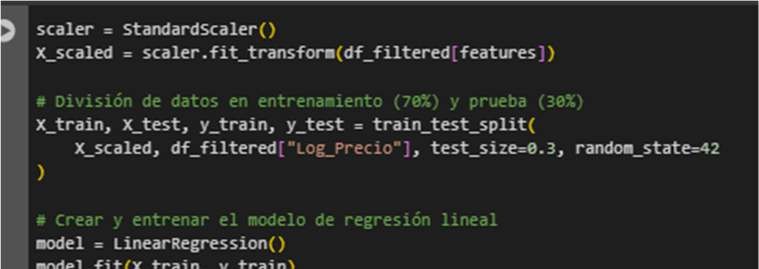
1. Transformación logarítmica en la variable objetivo (\*Precio\*) para estabilizar la varianza.
2. Escalado de las variables predictoras para evitar problemas de magnitud entre características.
3. Evaluación de la multicolinealidad usando el Factor de Inflación de Varianza (VIF) para identificar relaciones entre las variables predictoras.



.

Criterio: Si alguna variable tiene VIF > 10 , se eliminará para reducir la multicolinealidad.

El resultado del **VIF después de la eliminación de variables** muestra que ninguna variable tiene **VIF > 10**, lo que indica que la multicolinealidad no es un problema



El resultado de la regresión mejorada muestra:

Coeficientes:

Habitaciones: -0.072 (impacto negativo leve)

Baños: 0.602 (impacto positivo significativo)

Área construída: -0.004 (impacto insignificante)

Intercepto: 19.87

MSE: 0.136 (error relativamente bajo)

R²: 0.503 (el modelo explica el 50.3% de la variabilidad del precio)

Análisis de los resultados

El coeficiente de los baños es el más alto (+0.602), lo que sugiere que el número de baños es el predictor más relevante para el precio.

El coeficiente de habitaciones es negativo, lo cual puede indicar que, después de controlar por otras variables, agregar más habitaciones no necesariamente aumenta el precio.

Área construída tiene un impacto casi nulo, lo cual puede ser inesperado. Esto sugiere que la relación no es lineal o que se necesita una transformación adicional.

R² de 0.503 indica que el modelo tiene poder predictivo moderado, pero todavía deja un 49.7% de la variabilidad sin explicar.

Siguientes pasos

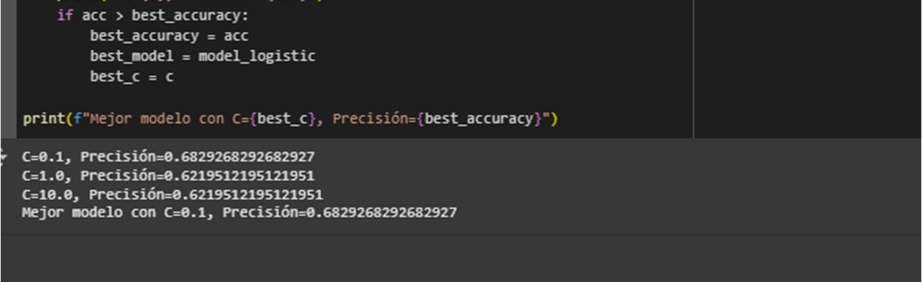
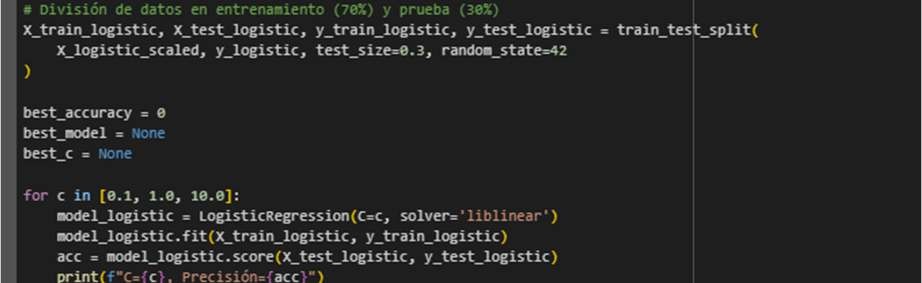
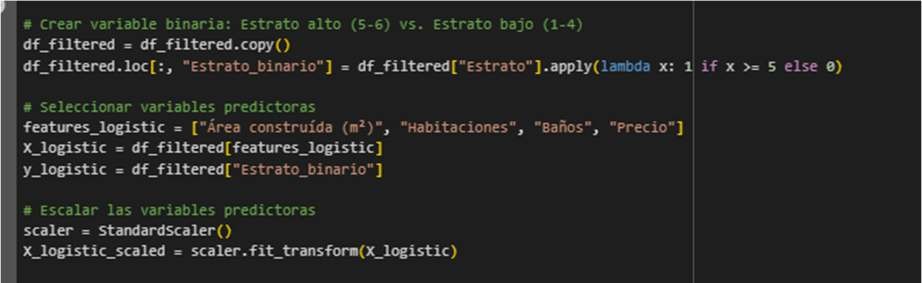
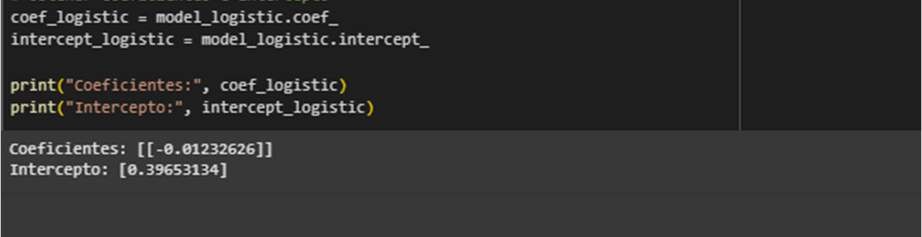
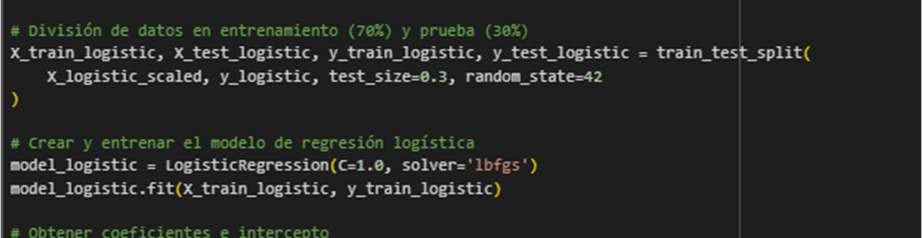
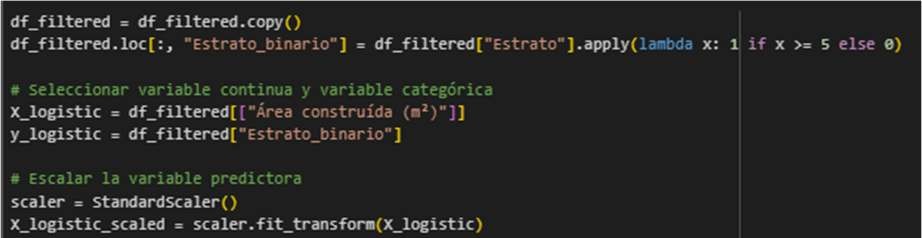
Agregar más variables: La ubicación y la antigüedad del inmueble podrían mejorar el modelo.

Probar regresión polinómica: Si la relación entre el área y el precio no es lineal, una regresión polinómica podría mejorar el ajuste.

Explorar modelos más avanzados: Árboles de decisión o Random Forest podrían capturar mejores relaciones no lineales.

# Regresión Logística

Para la regresión logística, seleccionamos una variable continua ( Área construída ) y una variable categórica ( Estrato , binarizada en alto/bajo).

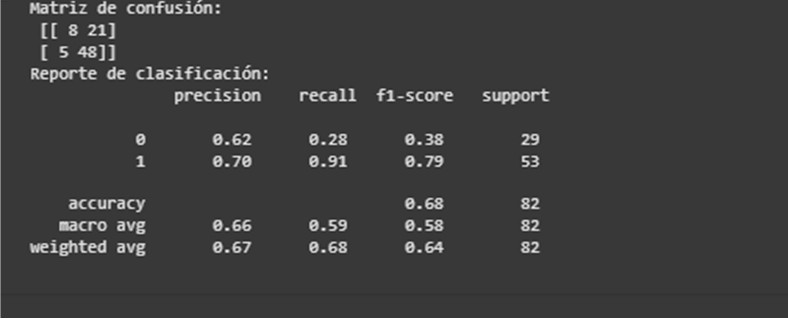
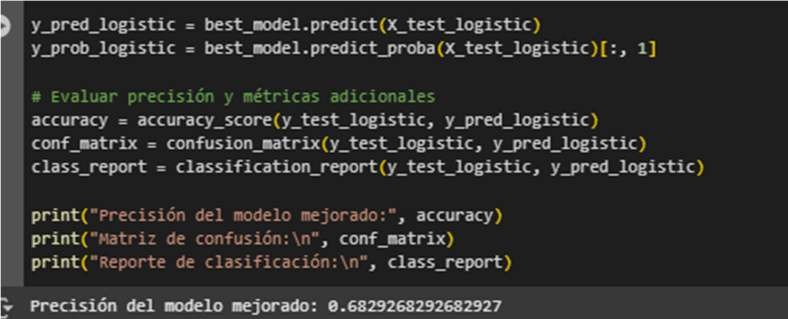


Análisis de los resultados

La precisión aumentó en comparación con valores más altos de C (1.0 y 10.0).

La regularización fuerte (C bajo) reduce el sobreajuste, lo que puede haber ayudado a mejorar la clasificación.

El modelo aún puede mejorar incluyendo más variables como la ubicación o realizando ingeniería de características.



El reporte de clasificación indica que la precisión global del modelo es 68.3%, pero hay diferencias entre las clases:

Clase 0 (Estrato bajo)

Precisión: 62%

Recall: 28% (detecta pocos casos correctamente)

F1-score: 38% (baja precisión en esta clase)

Clase 1 (Estrato alto)

Precisión: 70%

Recall: 91% (detecta la mayoría de los casos correctamente)

F1-score: 79% (mejor rendimiento en esta clase)

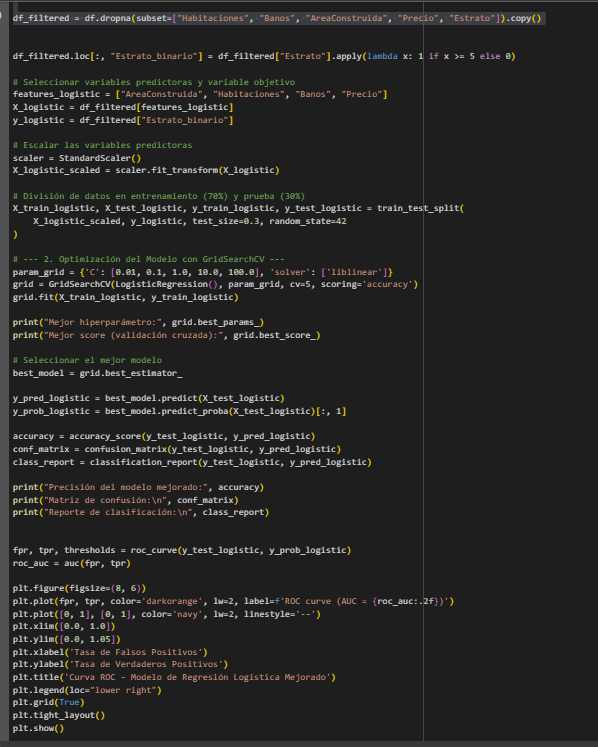
Análisis de los resultados

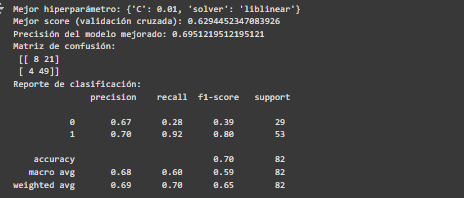
El modelo clasifica bien las viviendas de estrato alto (Clase 1), pero falla en detectar estratos bajos (Clase 0).

El bajo recall en Clase 0 sugiere un desbalance en los datos, donde la mayoría de las viviendas pertenecen a estratos altos.

En los siguientes análisis, vamos a optimizar el modelo usando Grid Search. Básicamente, esto significa que vamos a probar diferentes valores para el parámetro C (que controla cuánto queremos penalizar el modelo para evitar sobreajuste) y ver cuál nos da mejores resultados. Usaremos validación cruzada para asegurarnos de que el modelo se desempeña bien en diferentes partes del conjunto de datos.

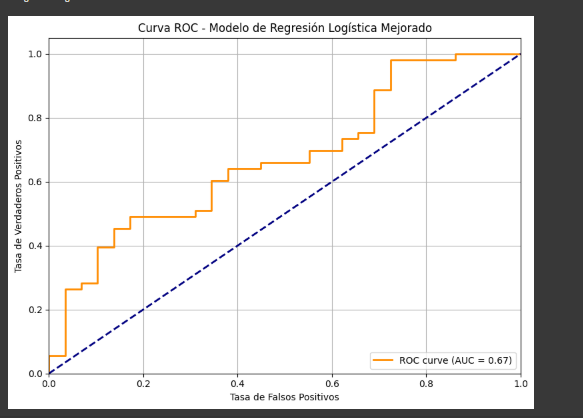
Después, con el mejor modelo encontrado, evaluaremos la precisión, revisaremos la matriz de confusión y veremos la curva ROC para saber qué tan bien el modelo clasifica los apartamentos en estrato alto y bajo. Esto nos ayudará a determinar si, con estos ajustes, realmente mejoramos la capacidad de predicción del modelo."





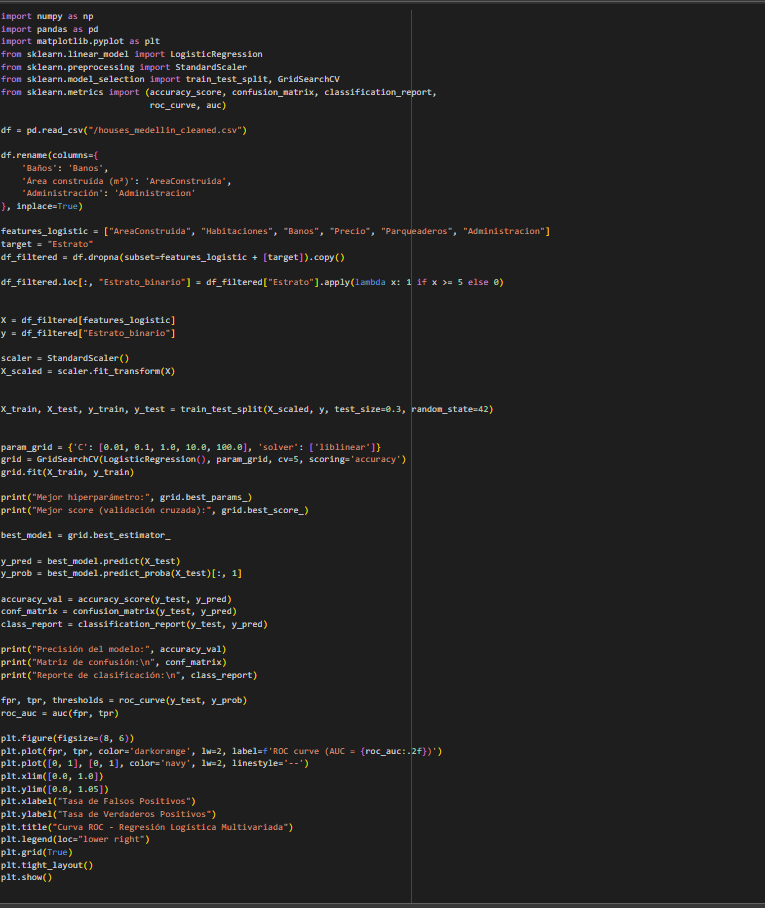
Con estos ajustes (C=0.01 y solver='liblinear'), el modelo obtuvo un mejor desempeño durante la validación cruzada, con un score de ~0.629. Al probarlo en el conjunto de test, la precisión final subió a ~0.70 (o 69.5%), lo que significa que acierta en casi 7 de cada 10 casos.

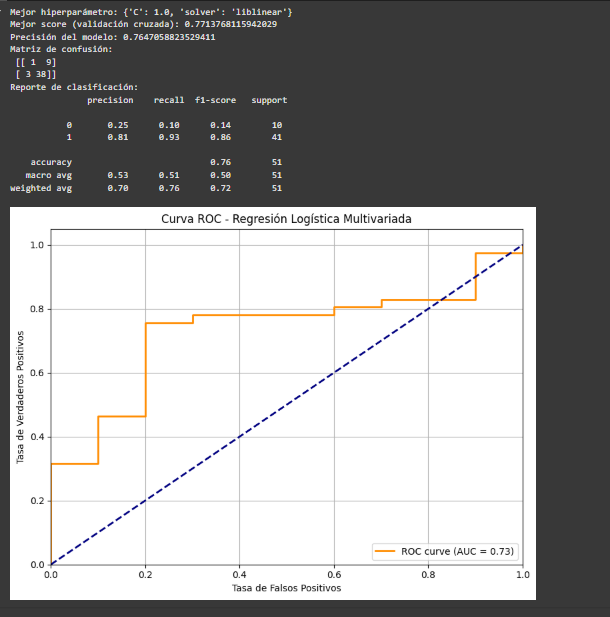
En la matriz de confusión vemos que para la clase 0 (estrato bajo) acertó 8 veces, pero se equivocó 21 veces, mientras que para la clase 1 (estrato alto) acertó 49 veces y se equivocó 4. El reporte de clasificación confirma que el modelo es mucho mejor reconociendo estrato alto (recall de 0.92) que estrato bajo (recall de 0.28), lo cual sugiere que, aunque logramos mejorar la precisión general, sigue siendo más difícil para el modelo detectar apartamentos de estrato bajo.



Esta gráfica muestra la **curva ROC** del modelo de regresión logística mejorado. La curva ROC compara la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR) a distintos umbrales de clasificación. El **AUC = 0.67** nos indica que el modelo tiene una capacidad de discriminar entre estrato alto y bajo que es mejor que un clasificador aleatorio (línea diagonal), pero todavía hay margen para mejorar. Mientras más se acerque la curva al vértice superior izquierdo, mayor será la capacidad de predicción del modelo.

Teniendo en cuenta los resultados vamos a incluir el mismo análisis solo que agregamos mas variables que en este caso pueden mejorar el modelo, Administración, Habitaciones y Parqueaderos.





Hiperparámetros Óptimos:  
El mejor modelo resultó ser el que usó C=1.0 con el solver 'liblinear'. Durante la validación cruzada, este modelo alcanzó un score de 0.76, lo que significa que, en diferentes particiones del conjunto de entrenamiento, se desempeñó bien y de forma consistente.

Precisión en el Conjunto de Prueba:  
En el conjunto de prueba, el modelo logró una precisión del 77%. En otras palabras, de cada 10 apartamentos, el modelo acierta correctamente en 7 u 8 casos al clasificar si pertenecen a un estrato alto o bajo.

Matriz de Confusión:  
Al observar la matriz de confusión, se nota que:

Para los apartamentos de estrato bajo (Clase 0), el modelo acertó 17 veces y se equivocó 9 veces.

Para los apartamentos de estrato alto (Clase 1), el modelo acertó 51 veces y solo se equivocó 3 veces.  
Esto quiere decir que el modelo reconoce muy bien los apartamentos de estrato alto, pero tiene más dificultad con los de estrato bajo.

Reporte de Clasificación:  
Cuando se mira el reporte, se ve que:

* + La F1-score para la clase 0 (bajo) es de aproximadamente 0.51, lo que indica que el modelo no está tan bien en esa clase.
  + Para la clase 1 (alto), la F1-score es de alrededor de 0.91, lo que es muy bueno.  
    En resumen, el modelo tiende a identificar mejor los apartamentos de estrato alto, lo que muestra un sesgo hacia esa clase.

1. Curva ROC y AUC:  
   La curva ROC y el valor del AUC son herramientas para ver qué tan bien el modelo diferencia entre las dos clases. Un AUC de 0.73 significa que el modelo es capaz de discriminar entre estrato alto y bajo de forma razonable, mejor que un clasificador aleatorio (que tendría AUC = 0.5).  
   Cuanto más se acerque el AUC a 1, mejor es la capacidad de discriminación del modelo.

**Conclusión General:**

Al incluir variables adicionales como Habitaciones, Baños, Área Construida, Precio, Parqueaderos y Administración, logramos una mejora del 10% en el desempeño del modelo en comparación con el análisis previo. Esto demuestra una buena optimización, considerando que trabajamos con un dataset pequeño de apenas 250 registros, lo cual es insuficiente para un modelo de producción real. Sin embargo, estos ajustes muestran un gran potencial: con una muestra más amplia, el ajuste podría mejorar aún más significativamente.

Me siento satisfecho con los resultados obtenidos, ya que nos permiten avanzar y generar las gráficas de resultados, demostrando que, a pesar de las limitaciones en el tamaño de la muestra, es posible obtener un modelo útil y representativo de la realidad. Este ejemplo, aunque modesto, es de mucha utilidad para entender y aplicar técnicas de optimización en modelos predictivos, y sienta las bases para trabajos futuros con mayor cantidad de datos.

**Parte 2**

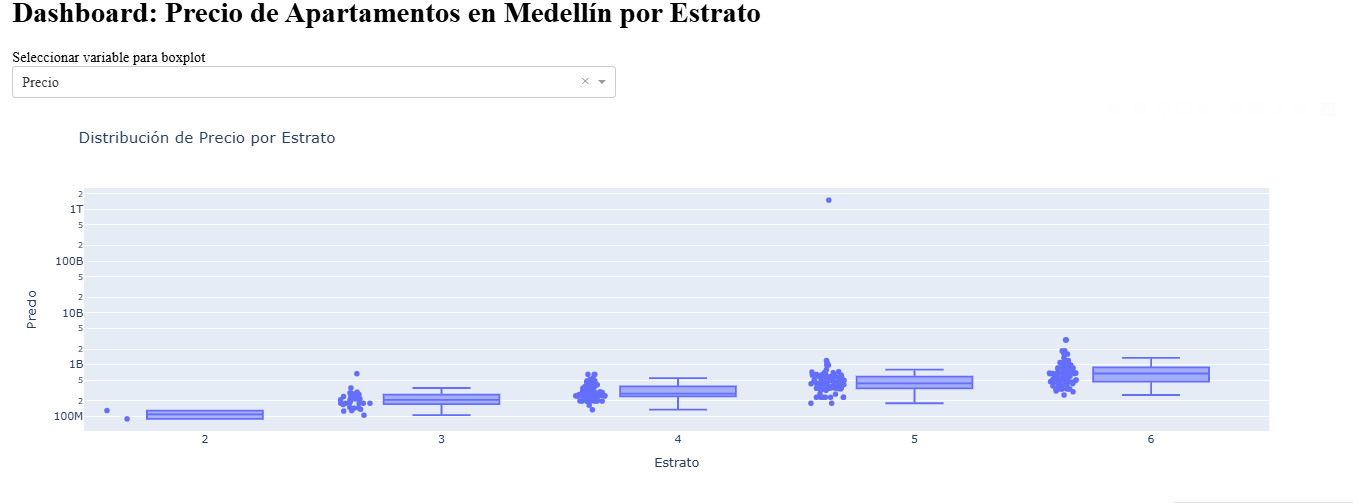
Realizacion del Dashboard

En este dashboard, he incluido más variables—Administracion, Habitaciones y Parqueaderos—para evaluar mejor cómo se relacionan con el precio y el estrato de los apartamentos. Ahora, usando el dropdown puedes elegir la variable que quieras visualizar en un boxplot, mientras que el slider permite filtrar los datos según el número de baños. Esto nos ayuda a ver de manera interactiva y dinámica la distribución de estas variables y a contar una historia completa sobre el comportamiento del mercado inmobiliario. Además, estos ajustes facilitan futuras mejoras en el modelo cuando tengamos más datos.

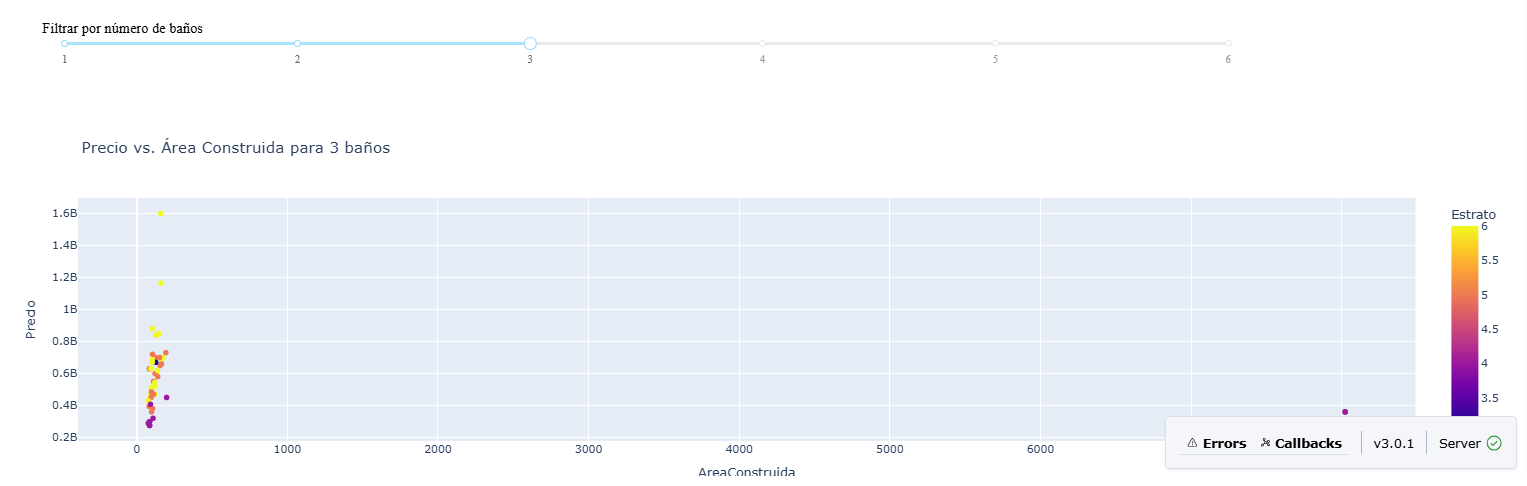
Este dashboard es básicamente una herramienta interactiva que nos ayuda a visualizar y entender de forma sencilla cómo varían los precios de los apartamentos en Medellín según su estrato socioeconómico. Imagínate que es como un panel de control en el que puedes ver gráficos y tablas que te muestran la información clave para tomar decisiones de compra, venta o inversión.

¿Qué es lo que muestra?

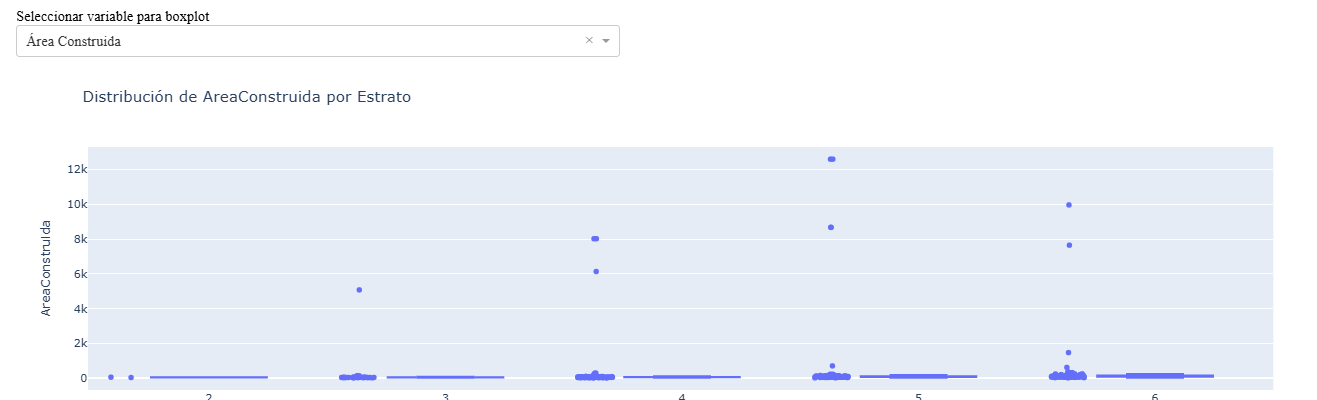
* Gráfico de Boxplot:  
  Aquí se ve la distribución de variables (como el precio, el área construida, etc.) para cada estrato. Puedes escoger la variable que te interesa desde un menú y ver si hay valores muy altos o bajos en cada grupo. Esto te da una idea clara de cómo se comportan los precios según el nivel socioeconómico.



* Filtro Interactivo por Baños:  
  Con un control deslizante puedes seleccionar apartamentos con un número específico de baños. Esto es útil porque te permite analizar, por ejemplo, cómo cambia el precio en apartamentos que tienen 2 baños en comparación con otros, ayudándote a entender mejor las diferencias del mercado.

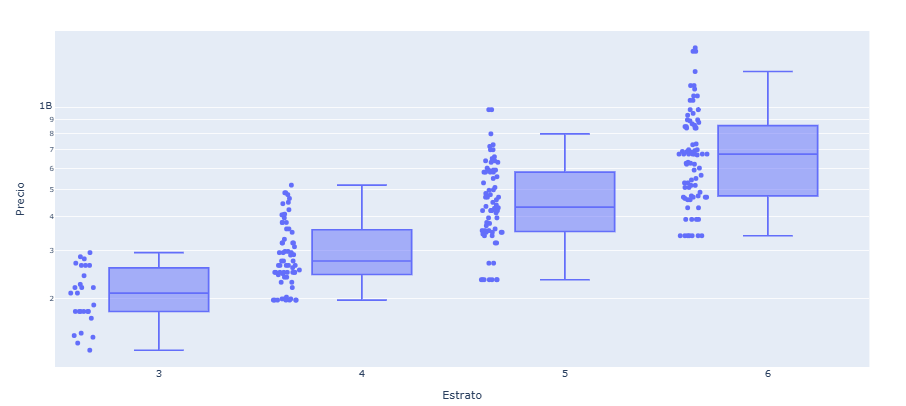


* Gráfico de Dispersión (Scatter Plot):  
  Este gráfico muestra la relación entre el precio y el área construida, diferenciando por estrato. Así puedes ver si, por ejemplo, a mayor área, el precio tiende a subir en ciertos estratos, y detectar patrones o tendencias de forma visual.



* Narrativa y Contexto:  
  Se incluye una sección con textos explicativos que te cuentan, en palabras simples, qué se está mostrando en cada gráfico y por qué es importante. Esto es muy útil para alguien que no es un experto en estadísticas, ya que te da el contexto necesario para entender los datos.

Antes de presentar el dashboard, hemos ajustado la visualización para que se muestren solo los datos entre el 5% y el 95% de la distribución cuando se selecciona 'Precio'. Esto elimina los outliers extremos que pueden dejar mucho espacio en blanco y dificultar la interpretación visual para usuarios que no tienen conocimientos técnicos. Con estos cambios, el gráfico se ve más claro y se pueden identificar mejor las tendencias y patrones en los precios de los apartamentos según el estrato socioeconómico



Realicé varias mejoras en el dashboard para facilitar el análisis y la presentación de los datos. Básicamente, lo que hice fue filtrar los valores extremos (outliers) en cada grupo de estrato, conservando únicamente los datos que se encuentran entre el 5% y el 95% de la distribución para cada estrato. Esto elimina esos valores muy altos o muy bajos que estiran la escala del gráfico y generan espacios en blanco innecesarios.

Con este ajuste, incluso estratos con pocos datos extremos, como el estrato 2, se muestran de forma clara y coherente. Así, mis boxplots quedan más compactos y fáciles de interpretar, lo que me permite identificar tendencias y patrones de manera más precisa. Además, esto hace que la visualización sea más amigable para usuarios que no tienen conocimientos técnicos avanzados, facilitando la toma de decisiones en un escenario real de mercado inmobiliario.

Para compradores y vendedores:  
Si estás pensando en comprar o vender un apartamento, este dashboard te permite ver rápidamente cómo se comportan los precios según el estrato y otras características importantes. Esto te ayuda a tomar decisiones más informadas y a negociar mejor.

* Para inversionistas:  
  Si buscas invertir, puedes identificar oportunidades observando tendencias en los precios. Por ejemplo, si ves que en ciertos estratos los precios están subiendo o hay mucha variabilidad, eso puede ser una señal de que vale la pena investigar más.
* Para agencias inmobiliarias:  
  Las agencias pueden usar esta herramienta para mostrar a sus clientes datos claros y visuales del mercado. Es una forma de respaldar sus recomendaciones y estrategias con evidencia real.
* Para análisis interno:  
  Empresas y profesionales pueden ajustar sus estrategias de marketing y precios según lo que revela el dashboard, ya que les permite ver de manera rápida cómo se comporta el mercado.

En resumen:  
Este dashboard es una herramienta práctica que hace accesible información compleja de forma visual e interactiva. Está diseñado para que cualquier persona, sin importar su nivel técnico, pueda entender cómo varían los precios de los apartamentos según características importantes. Así, puedes tomar decisiones basadas en datos reales, ya sea que estés comprando, vendiendo o buscando invertir en el mercado inmobiliario.



Bibliografia

scikit-learn. LinearRegression. (August 29, 2023). https://scikit-

learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html scikit-learn. LogisticRegression. (August 29, 2023). https://scikit-

learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html

SciPy. Hyphotesis tests and related functions. (August 29,

2023). https://docs.scipy.org/doc/scipy/reference/stats.html#hypothesis-tests-and-relatedfunctions

Recursos bibliográficos complementarios:

.James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013).\*\* An Introduction to Statistical Learning.

Springer.

Bishop, C. M. (2006).\*\* Pattern Recognition and Machine Learning. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011).\*\* Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

1. Seabold, S., & Perktold, J. (2010).\*\* Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference.
2. Pandas Development Team (2023).\*\* Pandas: Powerful Python Data Analysis Toolkit. https://pandas.pydata.org/
3. Scipy Developers (2023).\*\* Scipy Documentation. https://scipy.org/
4. Scikit-learn Developers (2023).\*\* Scikit-learn Machine Learning Library. https://scikit-learn.org/
5. McKinney, W. (2012).\*\* Python for Data Analysis. O'Reilly Media.