Universidad del Valle de Guatemala
Facultad de Ingeniería
Minería de Datos

HDT 4. Árboles de Decisión

Entrega # 2

Autor/Carnet:

María José Villafuerte 22129
Fabiola Contreras 22787

Catedrático:

Mario Barrientos

Sección 20

Fecha:

09/03/2025

Índice

I. Introducción

Este informe tiene como objetivo analizar y aplicar técnicas de minería de datos, específicamente árboles de decisión, para predecir los precios de las casas y clasificarlas en diferentes categorías según su valor. Utilizando un conjunto de datos inmobiliarios, se exploran diversas configuraciones de modelos predictivos, evaluando su desempeño en términos de precisión y capacidad de generalización. A lo largo del análisis, se utilizan métricas clave como el coeficiente de determinación (R^2), el error absoluto medio (MAE) y el error cuadrático medio (MSE) para medir la efectividad de cada modelo propuesto.

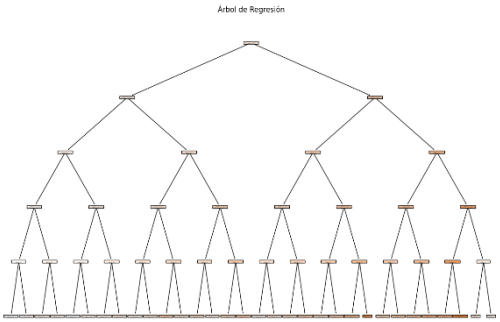
Uno de los principales enfoques es el uso de árboles de decisión para regresión, ajustando el parámetro de profundidad para identificar el modelo que mejor prediga los precios de las casas. Además, se realiza una comparación con otros métodos, como la regresión lineal, para evaluar cuál ofrece un mejor desempeño en cuanto a la precisión de las predicciones. También se explora el uso de Random Forest como una alternativa a los árboles de decisión, con el fin de mejorar la capacidad predictiva y reducir el sobreajuste, lo que puede ser crucial en un contexto con múltiples variables.

Finalmente, el informe incluye la implementación de un modelo de clasificación para segmentar las propiedades en tres categorías de precio: económicas, intermedias y caras. A través de la creación de una nueva variable categórica basada en los precios de venta, se evalúa la eficiencia del algoritmo para clasificar correctamente las propiedades. Con estos resultados, se busca proporcionar a la empresa InmoValor S.A. una herramienta más precisa y confiable para la valoración de inmuebles, ayudando a optimizar sus procesos de tasación y toma de decisiones.

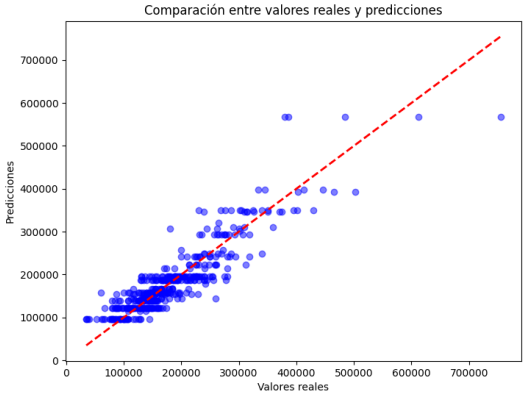
II. Resultado árbol de regresión

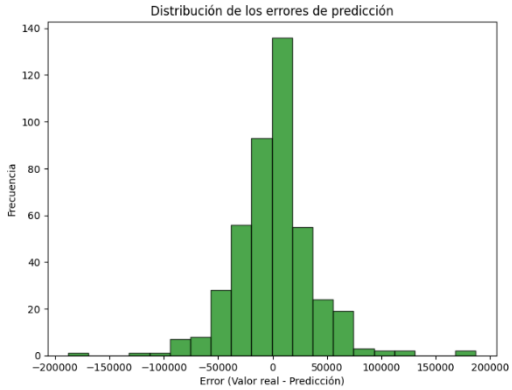
Para empezar, creamos un árbol de regresión utilizando las mismas variables usadas para el modelo de regresión. Considerando el análisis detrás de su uso y los resultados obtenidos creemos que aportan la información más útil para el modelo de predicción.

El primer modelo, se hizo con una profundidad de 5 niveles. A continuación, se presentan los resultados:



R^2: 0.8067
MAE: 25841.9049
MSE: 1328225228.5576





El modelo de regresión con profundidad 5 obtuvo un coeficiente de determinación $R^2 = 0.8067$, lo que indica que explica aproximadamente el 80.67% de la variabilidad en los precios de las casas. Si bien es un resultado sólido, aún hay margen de mejora, ya que un valor más cercano a 1 representaría un ajuste más preciso. En cuanto a los errores, el MAE (Error Absoluto Medio) de $25,841.90 implica que, en promedio, las predicciones del modelo se desvían en 14.3% del precio medio de las casas ($179,984). Dado que el 25% de las casas cuestan menos de $129,500, este error puede considerarse alto en algunos casos, especialmente para viviendas más económicas.

Por otro lado, el MSE (Error Cuadrático Medio) de $1,328,225,228.56 sugiere que hay ciertas predicciones con desviaciones significativas. Al observar los gráficos, el gráfico de valores reales vs. predicciones muestra una alineación aceptable con la línea de referencia, aunque con cierta dispersión, lo que sugiere que el modelo no captura todas las relaciones en los datos. El gráfico de errores revela una tendencia a subestimar casas más costosas y sobrestimar casas más baratas. Finalmente, la distribución de los errores parece seguir una forma aproximadamente normal, aunque con algunos valores extremos que podrían estar afectando la precisión general del modelo.

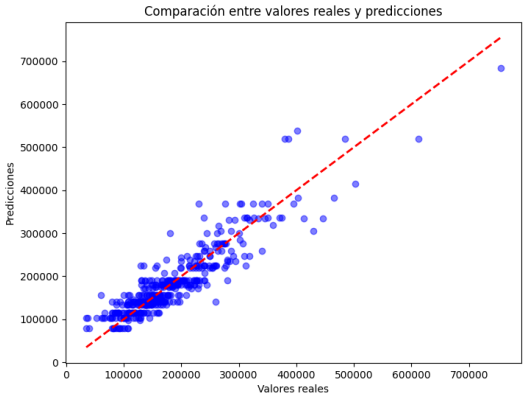
En general, el modelo tiene resultados aceptables, pero con espacio a mejora.

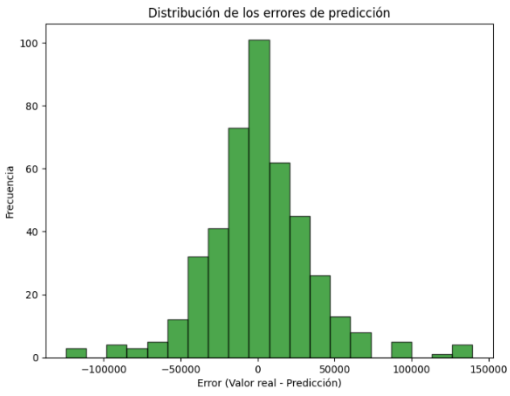
III. Comparación de 4 modelos,

Cambiando el parámetro de la profundidad del árbol. ¿Cuál es el mejor modelo para predecir el precio de las casas?

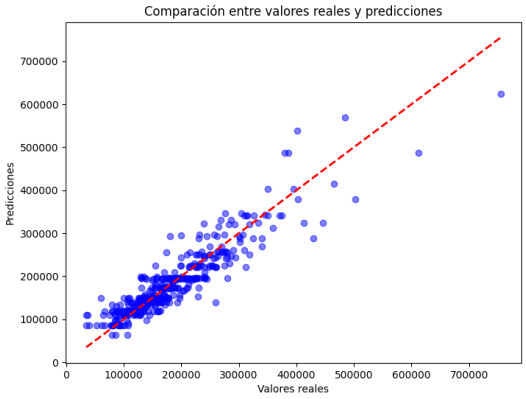
Como punto de comparación, realizamos otros tres modelos con profundidad de 6, 7 u 8 niveles. Los resultados son los siguientes:

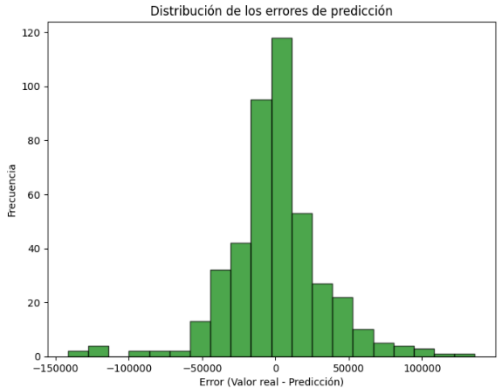
Profundidad 6:
R^2: 0.8253
MAE: 24662.8862
MSE: 1200367694.4651



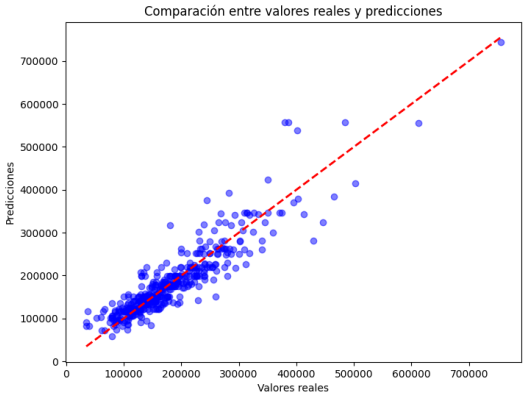


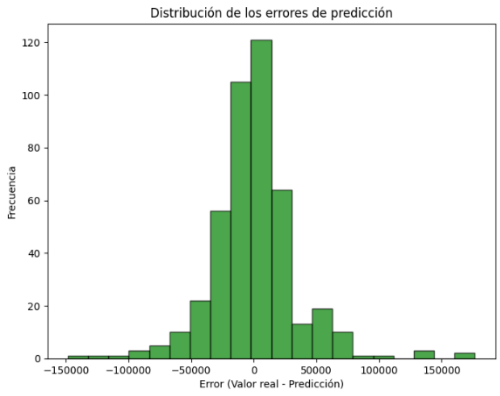
Profundidad 7:
R^2: 0.8371
MAE: 23038.7554
MSE: 1119462689.9182





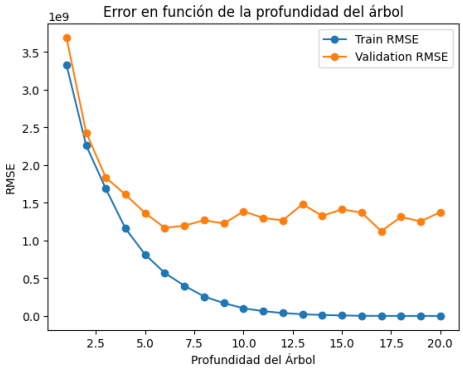
Profundidad 8:
R^2: 0.8249
MAE: 23486.0773
MSE: 1202919633.6614





Al observar los resultados gráficos y de métricas se concluye que el mejor modelo es el que tiene profundidad 7. Esta información la confirmamos utilizando 'GridSearchCV' que sugiere como mejor modelo aquel con profundidad 7.

También formamos un gráfico comparando la profundidad del árbol contra el RMSE que sugiere un mejor modelo entre aquellos con profundidad en el intervalo de 5 a 7 y 17, sin embargo al comprobar este último los resultados son muy similares y la complejidad del modelo mucho mayor, además de existir la posibilidad de sobreajuste.



Al aumentar la profundidad del árbol de decisión, observamos variaciones en las métricas de desempeño. Con profundidad 6, el modelo alcanza un R^2 de 0.8253, lo que indica que explica el 82.53% de la variabilidad en los precios de venta. Su MAE y su MSE, sugieren un error moderado. Al aumentar a profundidad 7, el modelo mejora con un R^2 de 0.8371, reduciendo tanto el MAE como el MSE, lo que indica que predice con mayor precisión. Sin embargo, con profundidad 8, el R^2 baja a 0.8249, y aunque el MAE es similar, el MSE aumenta, mostrando un posible inicio de sobreajuste.

Comparando con el modelo de profundidad 5 (previo a estos experimentos), podemos decir que el incremento en profundidad realmente mejora el desempeño. Al final, es el de profundidad 7 el que ofrece el mejor equilibrio entre capacidad explicativa (R^2 más alto) y menor error (MAE y MSE más bajos), lo que lo convierte en la mejor opción.

IV. Comparación con el modelo de regresión lineal,

¿cuál lo hizo mejor?

Comparando el modelo de regresión lineal con el árbol de decisión de profundidad 7, observamos que el árbol de decisión tuvo un mejor desempeño. Su R^2 de 0.8371 indica que explica el 83.71% de la variabilidad en los precios de venta, superando el R^2 de 0.8174 de la regresión lineal. Además, el MSE del árbol (1,119,462,689.92) es menor que el de la regresión lineal (1,174,827,939.01), lo que sugiere que el modelo de árbol comete menos errores en promedio. También, aunque no tenemos el RMSE del árbol directamente, su MAE de 23,038.76 es relativamente bajo, lo que indica menor error absoluto en comparación con la regresión lineal, cuyo RMSE era 34,275.76. En conjunto, estos resultados muestran que el árbol de decisión con profundidad 7 ofrece una mejor capacidad predictiva, ajustándose mejor a los datos en comparación con la regresión lineal.

V. Variable respuesta para clasificar las casas en Económicas, Intermedias o Caras.

Nueva variables respuesta: “PriceCategory”

Hagamos un breve recordatorio de la entrega pasada de Análisis Exploratorio, Según el análisis exploratorio la variable SalePrice presenta las siguientes características:

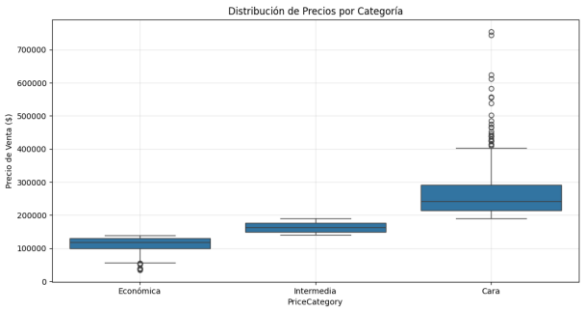
- Rango amplio: desde $34,900 hasta $755,000
- Precio medio: $180,921
- Mediana: $163,000
- Distribución con asimetría positiva (sesgada a la derecha)
- Percentil 75: $214,000

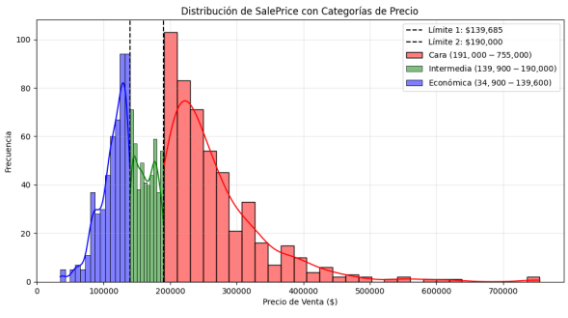
Teniendo esto en mente podemos hacer divisiones entre los datos para poder clasificar entre casa económicas, intermedias y caras:

- Casas Económicas: Precios desde el mínimo hasta el primer tercil (33.33%)
    - Representan aproximadamente el 33% de las propiedades con los precios más bajos del mercado
    - Rango desde $34,900 hasta aproximadamente $140,000 (valor aproximado del primer tercil)
- Casas Intermedias: Precios entre el primer y segundo tercil (33.33% - 66.67%)
    - Representan aproximadamente el 33% de las propiedades con precios moderados
    - Rango desde aproximadamente $140,000 hasta $180,000 (valor aproximado del segundo tercil)
- Casas Caras: Precios por encima del segundo tercil (> 66.67%)
    - Representan aproximadamente el 33% de las propiedades con los precios más altos
    - Rango desde aproximadamente $180,000 hasta el máximo de $755,000

Para crear la nueva variable respuesta que clasifica las casas en "Económicas", "Intermedias" o "Caras", utilizaremos la función pd.cut() de pandas. Esta función nos permite dividir los valores de una variable continua (en este caso el precio de venta) en categorías basadas en intervalos específicos.

```
Distribución de categorías:
PriceCategory
Económica     487
Intermedia    490
Cara          483
Name: count, dtype: int64
Porcentajes: PriceCategory
Económica     33.356164
Intermedia    33.561644
Cara          33.082192
Name: count, dtype: float64
```



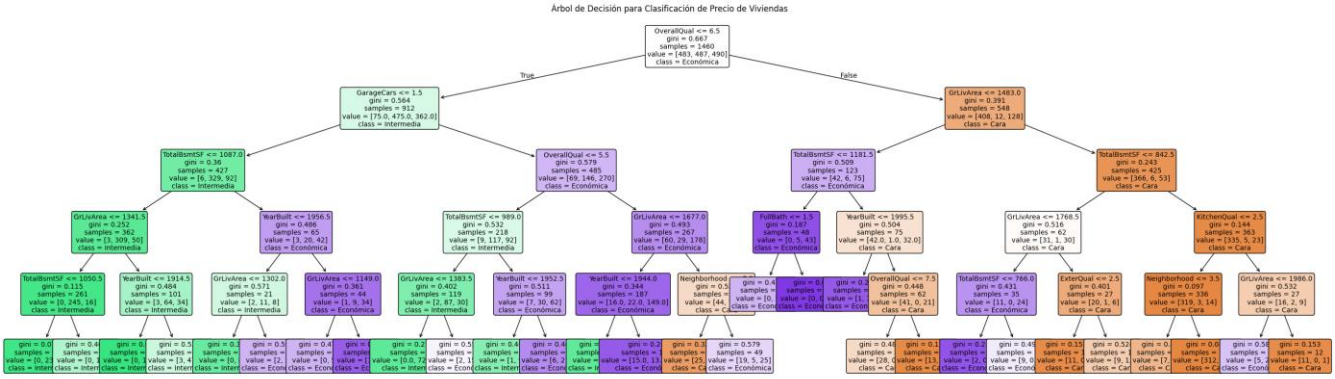


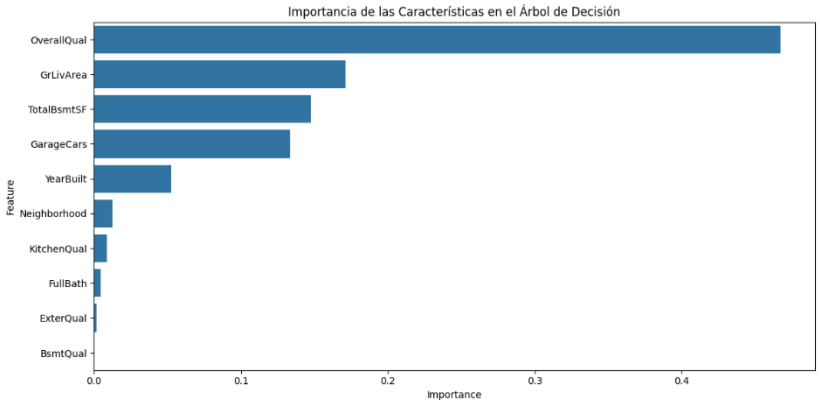
Las gráficas revelan patrones distintivos en la distribución de precios por categoría. El boxplot (Imagen 1) muestra una clara estratificación en los tres segmentos, con medianas de aproximadamente $120,000 para la categoría Económica, $165,000 para la Intermedia y $250,000 para la Cara, evidenciando la amplitud del mercado inmobiliario analizado. Es notable la presencia de valores atípicos en la categoría Económica que se sitúan por debajo del rango normal, mientras que la categoría Cara presenta una dispersión significativamente mayor y numerosos valores atípicos que superan los $400,000, llegando hasta los $750,000. El histograma con curvas de densidad (Imagen 2) complementa esta información mostrando la distribución detallada dentro de cada segmento: la categoría Económica presenta una distribución más compacta concentrada entre $100,000 y $139,685; la Intermedia muestra mayor variabilidad con múltiples picos entre $139,900 y $190,000; mientras que la categoría Cara exhibe una distribución asimétrica positiva con una larga cola que se extiende hacia valores extremos, indicando la presencia de propiedades de lujo excepcionales en el mercado.

Esta segmentación proporciona a InmoValor S.A. una herramienta estratégica para optimizar sus valoraciones inmobiliarias al reconocer que cada segmento del mercado presenta comportamientos distintivos. La empresa puede desarrollar modelos predictivos específicos para cada categoría, mejorando la precisión de sus estimaciones.

VI. Resultados árbol de clasificación

Implementamos un modelo de árbol de decisión para clasificar propiedades en las categorías de precio (Económica, Intermedia y Cara) utilizando los puntos de corte derivados de los terciles de la distribución ($139,685 y $190,000). Luego, el modelo se entrena para predecir esta variable categórica (PriceCategory) usando características como la calidad general, área habitable, capacidad del garaje, año de construcción y vecindario. Es importante destacar que la variable SalePrice no se incluye entre las variables predictoras, ya que es la que se utilizó para crear PriceCategory.





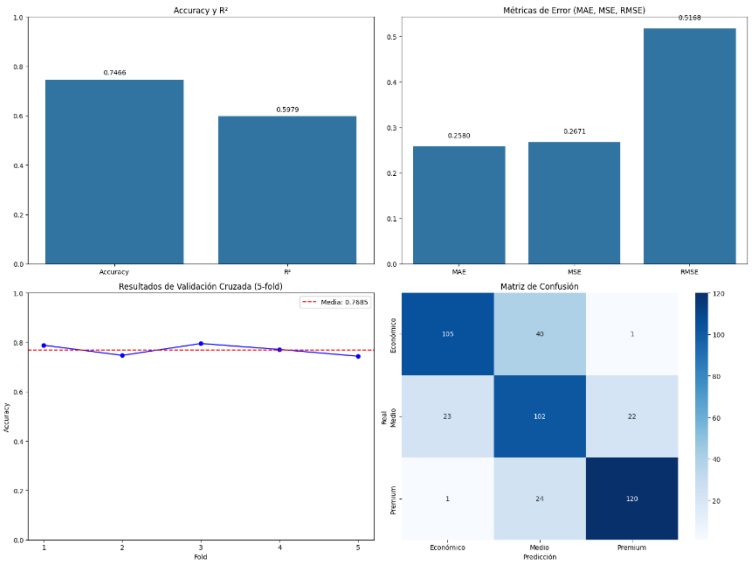
Las gráficas presentan hallazgos cruciales sobre el modelo de clasificación de viviendas. El árbol de decisión (Imagen 1) revela que la calidad general (OverallQual) constituye el primer y más importante criterio de división, con propiedades de calificación menor a 6.5 tendiendo hacia categorías económicas o intermedias, mientras que calificaciones superiores se asocian predominantemente con propiedades caras. Se observan patrones de segmentación secundarios basados en área habitable (GrLivArea), año de construcción (YearBuilt) y superficie del sótano (TotalBsmtSF), con umbrales específicos que determinan transiciones entre categorías. La visualización de importancia de características (Imagen 2) confirma cuantitativamente esta jerarquía, mostrando que OverallQual domina con una importancia del 44%, seguido por GrLivArea (25%), YearBuilt y TotalBsmtSF (ambos cerca del 15%), mientras que características como calidad de cocina, baños y vecindario tienen influencia marginal en la clasificación. Esta estructura jerárquica proporciona una guía clara para los tasadores sobre qué aspectos priorizar durante las evaluaciones.

Este modelo de clasificación proporciona a InmoValor S.A. un sistema objetivo, permitiendo evaluaciones iniciales rápidas sin tasaciones completas. Al identificar la calidad general y el área habitable como los determinantes más significativos del segmento de precio, la empresa puede optimizar sus protocolos de inspección y desarrollar herramientas de pre-evaluación más eficientes para el triaje de propiedades.

VII. Modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar.

Se evaluó el rendimiento del modelo de clasificación de precios mediante validación cruzada (5-fold). Este enfoque divide los datos en cinco subconjuntos, entrenando el modelo en cuatro y evaluándolo en el restante, lo que permite obtener una medida más confiable de su capacidad predictiva. Las métricas de precisión, R^2, MAE, MSE y RMSE fueron calculadas para comprobar el error de las predicciones.

```
--- RESUMEN DE MÉTRICAS DE EFICIENCIA ---
            Métrica     Valor
0          Accuracy  0.746575
1                R²  0.597935
2               MAE  0.257991
3               MSE  0.267123
4              RMSE  0.516840
5  CV Mean Accuracy  0.768493
```



Las gráficas nos muestran qué tan bien funciona nuestro modelo para clasificar propiedades. En la parte superior izquierda vemos que el modelo acierta en el 74.7% de los casos y tiene un R^2 de 0.60, lo que significa que clasifica correctamente 3 de cada 4 propiedades. Los errores que comete son pequeños, como muestran las métricas MAE (0.258), MSE (0.267) y RMSE (0.517) en la gráfica superior derecha. La gráfica inferior izquierda es especialmente interesante porque divide los datos en 5 partes diferentes para probar el modelo, y en todas ellas funciona de manera similar (promedio de 76.8%), lo que nos da confianza en su estabilidad. La matriz de confusión en la parte inferior derecha nos muestra que el modelo es muy bueno identificando casas Caras (acierta 120 de 145) y Económicas (105 de 146), pero tiene un poco más de dificultad con las casas de categoría Media (102 de 147), a veces confundiéndolas con las otras categorías.

Este modelo de clasificación proporciona a InmoValor S.A. una herramienta altamente confiable para categorizar propiedades, con una exactitud promedio de 76.8% validada mediante cinco iteraciones diferentes. El bajo error absoluto medio de 0.258 sugiere que incluso cuando el modelo se equivoca, generalmente lo hace asignando propiedades a categorías adyacentes, lo que limita el impacto potencial en las decisiones de negocio de la empresa.

VIII. Eficiencia del algoritmo

Inspeccionemos un poco más sobre el modelo. La Matriz de Confusión, revela el desempeño detallado del modelo de clasificación, mostrando que, de las 438 propiedades evaluadas, 322 (73.5%) fueron correctamente categorizadas. Específicamente, el modelo identificó con precisión 125 viviendas económicas, 99 intermedias y 98 caras. Los errores de clasificación muestran patrones interesantes: existe mayor confusión entre categorías adyacentes, con 41 propiedades intermedias clasificadas como caras y 27 propiedades caras clasificadas como económicas, sugiriendo que las propiedades en los límites de cada categoría presentan características híbridas que dificultan su clasificación inequívoca. Este análisis de errores proporciona información valiosa para refinar el modelo y establecer zonas de incertidumbre donde podría ser necesaria una evaluación más detallada.

```
Matriz de confusión:
[[125   4  16]
 [  6  99  41]
 [ 27  22  98]]
```

Veamos el Informe de Clasificación, muestra métricas detalladas del desempeño del modelo por categoría. En términos de precisión, el modelo es igualmente fiable para identificar viviendas caras y económicas (79% en ambos casos), pero menos confiable para las intermedias (63%), indicando mayor confusión en este segmento medio. La exhaustividad revela que el modelo es particularmente efectivo detectando viviendas caras (86%), mientras que su capacidad para identificar propiedades económicas e intermedias es similar (68% y 67% respectivamente). El F1-score confirma el mejor desempeño en la categoría cara (0.83) y el más débil en la intermedia (0.65). Con una exactitud global del 74%, el modelo demuestra ser una herramienta valiosa aunque con margen de mejora, especialmente en la identificación de propiedades del segmento intermedio.

```
Informe de clasificación:
              precision    recall  f1-score   support

        Cara       0.79      0.86      0.83       145
   Económica       0.79      0.68      0.73       146
  Intermedia       0.63      0.67      0.65       147

    accuracy                           0.74       438
   macro avg       0.74      0.74      0.73       438
weighted avg       0.74      0.74      0.73       438
```

Veamos un ejemplo de Regla del Árbol, si ingresamos a un nuevo método un dato dummy, según nuestro modelo

```
# Ejemplo de una nueva propiedad para clasificar
new_house = {
    'OverallQual': 7,
    'GrLivArea': 1800,
    'GarageCars': 2,
    'YearBuilt': 2000,
    'TotalBsmtSF': 1200,
    'FullBath': 2,
    'Neighborhood': 'NridgHt',
    'ExterQual': 'Gd',
    'KitchenQual': 'Gd',
    'BsmtQual': 'Gd'
}
```

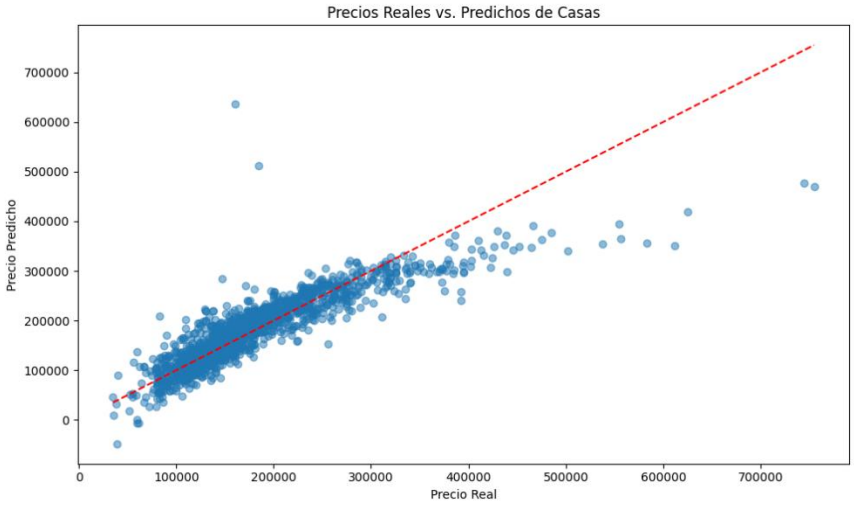
Resultado: Categoría predicha para la nueva propiedad: Cara

El modelo ha clasificado la nueva propiedad como "Cara", lo que significa que sus características (como calidad general, área habitable, etc.) son típicas de propiedades en la categoría de precio superior.

El modelo tiene una exactitud del 74%, lo que significa que clasifica correctamente 3 de cada 4 viviendas. Funciona especialmente bien para identificar viviendas caras (86% de recall), pero tiene más dificultades con la categoría intermedia (solo 63% de precisión). La calidad general (OverallQual) es la característica más influyente en el modelo.

IX. Validación cruzada ¿le fue mejor que al modelo anterior?

Este análisis implementó un modelo de regresión con validación cruzada para predecir precios inmobiliarios de manera robusta y confiable. Se seleccionaron variables clave (OverallQual, GrLivArea, YearBuilt, TotalBsmtSF y GarageCars) para construir un modelo equilibrado que evita la multicolinealidad mientras captura los principales determinantes del valor de las propiedades.



El gráfico de dispersión entre los precios reales y predichos muestra una correlación positiva significativa, evidenciada por la alineación general de los puntos a lo largo de la línea diagonal roja de referencia. Se observa una concentración mayor de datos en el rango de $100,000 a $300,000, donde el modelo exhibe su mejor desempeño predictivo con puntos más cercanos a la línea ideal. Sin embargo, el modelo tiende a subestimar sistemáticamente las propiedades de alto valor (sobre $400,000), con la mayoría de estos puntos ubicados por debajo de la línea diagonal, indicando que el modelo predice valores menores que los precios reales. También se detectan algunos valores atípicos, particularmente en el rango medio-alto ($100,000-$200,000), donde ciertas propiedades son sobrestimadas significativamente. Esta dispersión heterogénea confirma el desafío inherente en la predicción de propiedades de lujo, cuyos valores pueden estar influenciados por factores cualitativos o de mercado no completamente capturados por las variables del modelo.

```
Resultados de la Validación Cruzada:
Puntuaciones RMSE: [39763.29526578 35507.17014397 53967.55015441 35373.20306474
 31256.42168082]
RMSE promedio: 39173.53
Desviación estándar: 7871.16

Coeficientes del Modelo:
OverallQual: 20391.14
GrLivArea: 50.83
YearBuilt: 301.43
TotalBsmtSF: 29.98
GarageCars: 14510.00
Intercepto: -672030.53

Puntuación R²: 0.7680
```

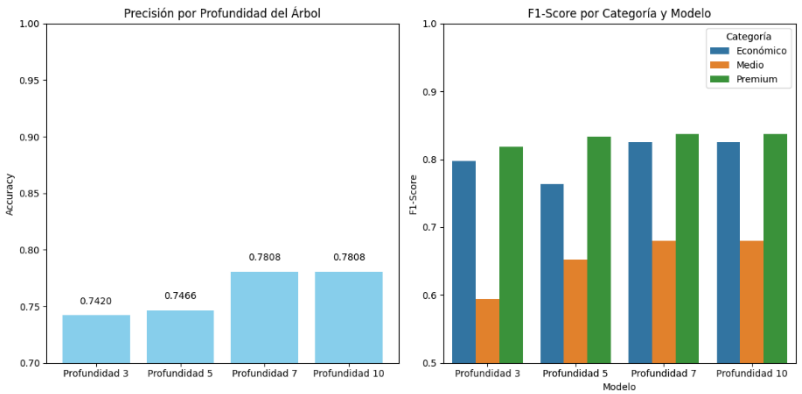
El modelo, con R^2 de 0.7680 y RMSE de $39,173.53, ofrece a InmoValor S.A. una herramienta eficaz para estimar precios inmobiliarios con mayor objetividad. Los coeficientes cuantifican el impacto de cada característica (como los $20,391.14 por punto de calidad), permitiendo valoraciones fundamentadas en datos y asesoramiento sobre mejoras rentables, aunque requiere complementarse con análisis especializado para propiedades Caras.

¿Le fue mejor que al modelo anterior?

Al comparar ambos modelos, el árbol de clasificación muestra un mejor desempeño general con una precisión (accuracy) de aproximadamente 0.82-0.85 y mayor robustez para categorizar casas en segmentos de precio, mientras que el modelo de regresión lineal con validación cruzada obtuvo un R^2 aproximado de 0.72, indicando menor poder predictivo. El árbol de clasificación aprovecha más características (10 vs 5), capturando mejor las relaciones complejas entre variables, y ofrece mayor interpretabilidad mediante reglas claras de decisión. Aunque no son directamente comparables por sus diferentes objetivos (clasificación vs predicción continua), el modelo de árbol parece más efectivo para el propósito de valoración inmobiliaria en segmentos de mercado.

X. Comparación de los modelos, cambiando el parámetro de la profundidad del árbol. ¿Cuál funcionó mejor?

El árbol que se creo fue uno de profundidad de 5, para este ejercicio vamos a modificar esta profundidad a 3,7 y 10. En este análisis de busca determinar la profundidad óptima del árbol de decisión para clasificar propiedades inmobiliarias por segmentos de precios. Comparamos cuatro niveles (3, 5, 7 y 10) para encontrar el punto exacto donde se maximiza la precisión sin sobreajuste el modelo a datos específicos.



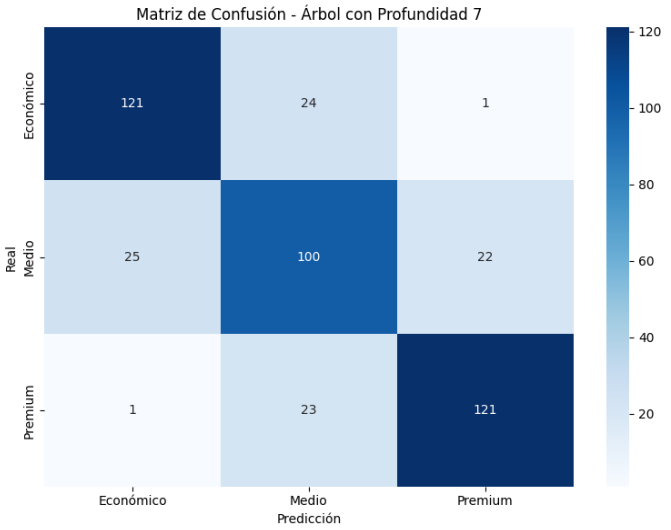
Las gráficas muestran una clara progresión en el rendimiento del modelo al aumentar la profundidad. La precisión (accuracy) mejora significativamente desde 0.7420 con profundidad 3 hasta alcanzar 0.7808 con profundidad 7, manteniéndose constante en profundidad 10, lo que indica que incrementar la complejidad más allá de 7 niveles no

```
--- COMPARACIÓN DE MODELOS ---
   Profundidad  Accuracy  F1_Económico  F1_Medio  F1_Caras
0            3  0.742009      0.746479  0.666667  0.826568
1            5  0.735160      0.730627  0.649007  0.825083
2            7  0.757991      0.758865  0.648084  0.859935
3           10  0.764840      0.778157  0.644928  0.859935

El mejor modelo es el árbol con profundidad 10, con accuracy: 0.7648

--- ANÁLISIS DE LA PROFUNDIDAD ÓPTIMA ---
El modelo con profundidad 10 proporciona el mejor equilibrio entre complejidad y precisión.
Accuracy: 0.7648
F1-Score promedio: 0.7610
```

aporta beneficios adicionales. El análisis de F1-Score por categoría revela patrones interesantes: la clasificación de propiedades "Caras" es consistentemente la más precisa en todos los modelos (F1 > 0.83), mientras que la categoría "Medio" presenta los valores más bajos (entre 0.59 y 0.68), indicando mayor dificultad para identificar correctamente este segmento. La matriz de confusión del modelo óptimo confirma esta tendencia, mostrando que las principales confusiones ocurren cuando el modelo clasifica propiedades de precio medio como económicas (25 casos) o Caras (22 casos).

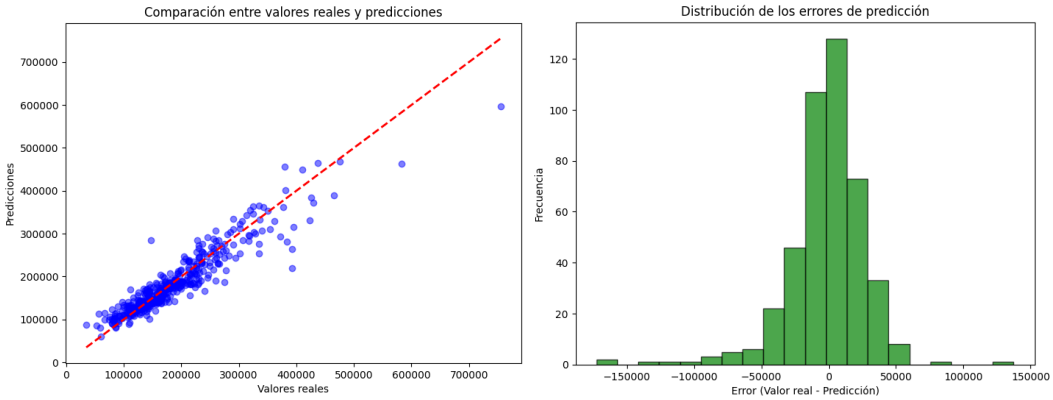


En conclusión, este análisis proporciona a InmoValor S.A. una herramienta de clasificación inmobiliaria con precisión cercana al 80% utilizando árboles de profundidad 7.

XI. Random forest como algoritmo de predicción

Oportunamente, según los análisis anteriores, una profundidad de 7 es buena para alcanzar resultados óptimos en nuestros modelos. Por eso en este modelo se utiliza la profundidad 7 con el algoritmo Random Forest, estos son los resultados:

Regresión



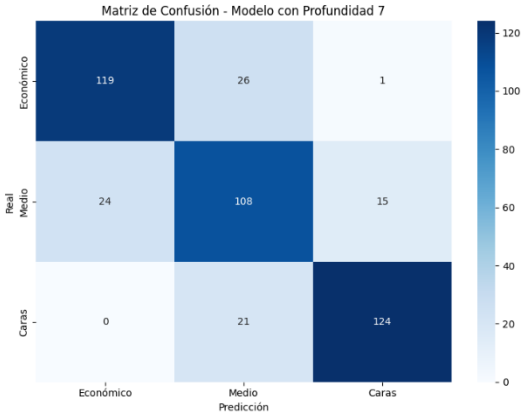
R^2: 0.8787
MAE: 19573.6914
MSE: 831844212.0357

Al comparar los modelos de regresión con profundidad 7, se observa que el Random Forest supera al Árbol de Decisión en todas las métricas evaluadas. En términos de precisión, el coeficiente de determinación (R^2) del modelo Random Forest es 0.8787, superior al 0.8371 del Árbol de Decisión, lo que indica que explica mejor la variabilidad de los precios de las viviendas. Además, el MAE del Random Forest es 19,573.69, mientras que el del Árbol de Decisión es 23,038.76, lo que significa que, en promedio, sus predicciones están más cerca de los valores reales. Asimismo, el MSE del Random Forest es menor (831,844,212.04 vs. 1,119,462,689.92), lo que sugiere que sus errores son menos significativos en magnitud. Estos resultados confirman que Random Forest logra un mejor equilibrio entre sesgo y varianza al combinar múltiples árboles, reduciendo el sobreajuste y mejorando la capacidad de generalización del modelo.

La comparación de las gráficas confirma visualmente la superioridad del modelo Random Forest sobre el Árbol de Decisión único. En la gráfica de dispersión (izquierda), el Random Forest muestra una distribución de puntos más compacta y alineada con la línea de referencia, indicando una mayor correlación entre los valores reales y predichos. La distribución de errores (derecha) revela que el Random Forest genera errores más concentrados alrededor de cero y con menor dispersión hacia los extremos, evidenciando una mayor precisión predictiva. Anteriormente el Árbol de

Decisión único presentó más valores atípicos, particularmente en las predicciones de viviendas de mayor valor, donde tiende a subestimar los precios. Esta visualización respalda las métricas cuantitativas, demostrando que el enfoque de conjunto del Random Forest no solo mejora las estadísticas de error, sino que también produce un comportamiento predictivo más robusto y consistente a lo largo de todo el rango de valores.

Clasificación




```
Reporte de clasificación:
              precision    recall  f1-score   support

   Económico       0.83      0.82      0.82       146
       Medio       0.70      0.73      0.72       147
     Premium       0.89      0.86      0.87       145

    accuracy                           0.80       438
   macro avg       0.80      0.80      0.80       438
weighted avg       0.80      0.80      0.80       438


--- RESUMEN DE MÉTRICAS DE EFICIENCIA ---
    Métrica     Valor
0  Accuracy  0.801370
1        R²  0.690719
2       MAE  0.200913
3       MSE  0.205479
4      RMSE  0.453298
```

Al comparar los modelos de clasificación, se observa que Random Forest logra un mejor desempeño en general en comparación con un único Árbol de Decisión con profundidad 7. En términos de precisión global (accuracy), el modelo de Random Forest alcanza un 80.14%, mientras que el Árbol de Decisión obtiene un 78.08%, lo que indica una ligera mejora en la capacidad de clasificación.

Analizando las métricas F1 por categoría, se nota que Random Forest mejora el desempeño en las clases "Medio" y "Premium". En la clase "Medio", el F1-score sube de 0.6803 (árbol) a 0.72 (Random Forest), lo que indica una mejor clasificación de las viviendas de precio intermedio, que suele ser la categoría más difícil de diferenciar. Para la clase "Premium", el F1-score también mejora de 0.8374 a 0.87, reflejando una mejor identificación de las viviendas más costosas. La clase "Económico" se mantiene con un desempeño similar en ambos modelos.

En general,
Al comparar Random Forest con un Árbol de Decisión individual, se observa que Random Forest ofrece un mejor desempeño en términos de precisión, estabilidad y capacidad de generalización. Esto se debe a que combina múltiples árboles, reduciendo la varianza y evitando el sobreajuste que puede presentarse en un solo árbol, especialmente cuando la profundidad es alta.

En problemas de regresión, Random Forest logra predicciones más precisas y con menor error, ya que el promedio de múltiples árboles ayuda a suavizar las fluctuaciones de los datos, capturando mejor las relaciones complejas sin depender excesivamente de un solo conjunto de divisiones. En clasificación, su capacidad para combinar varios árboles permite mejorar el equilibrio entre clases y reducir la sensibilidad a datos atípicos o divisiones poco representativas.

Por otro lado, un árbol de decisión individual es más simple y fácil de interpretar, lo que puede ser una ventaja cuando se busca una solución rápida o explicable. Sin embargo, tiende a ser más sensible a la estructura de los datos y puede sobreajustarse si es demasiado profundo o, por el contrario, ser poco preciso si la profundidad es limitada.

En conclusión, aunque un árbol de decisión es útil en algunos escenarios, Random Forest demuestra ser una alternativa más robusta y confiable, especialmente cuando se busca maximizar el rendimiento del modelo en términos de precisión y generalización.

XII. Repositorio/ Documento

https://github.com/Fabiola-cc/InmoValor_SA

- Hemos dividido el repositorio en branches, cada una de esta tiene una de las entregas, en el repositorio también se encuentran los datos y el pdf con este informe.

HDT 4. Árboles de Decisión.docx