title: "proyecto2"

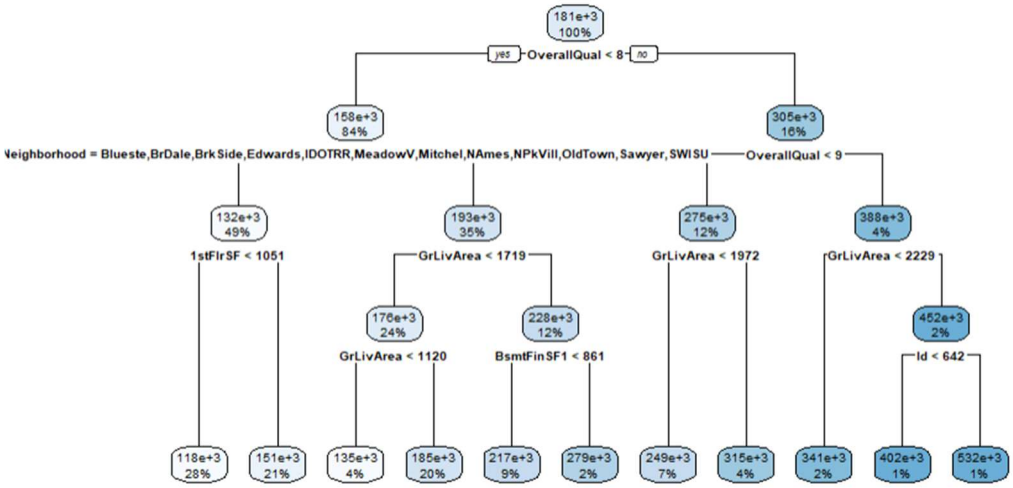
date: "2025-01-30"

Usando datos

esta vez se usarán todos los datos

haciendo el Árbol

A continuación vamos a intentar hacer un árbol y intentando hacerlo de manera en que muestre patrones que previamente puede que fuesen omitidas



Resultados

La variable más importante es OverallQual.

Si OverallQual < 8, el precio promedio baja a 158,000 USD.

Si OverallQual ≥ 8, el precio promedio sube a 305,000 USD.

Para viviendas con OverallQual < 8 (84% de los datos):

El barrio (Neighborhood) influye bastante. Si pertenece a ciertos barrios de menor precio, el valor disminuye.

1stFlrSF < 1051 pies cuadrados indica precios bajos (~118,000 USD).

GrLivArea (área habitable sobre el suelo) también es clave. Si es menor a 1,120, el precio baja (~135,000 USD), pero si es mayor, sube (~185,000 USD).

BsmtFinSF1 (área terminada del sótano) también juega un rol en el precio.

Para viviendas con OverallQual ≥ 8 (16% de los datos):

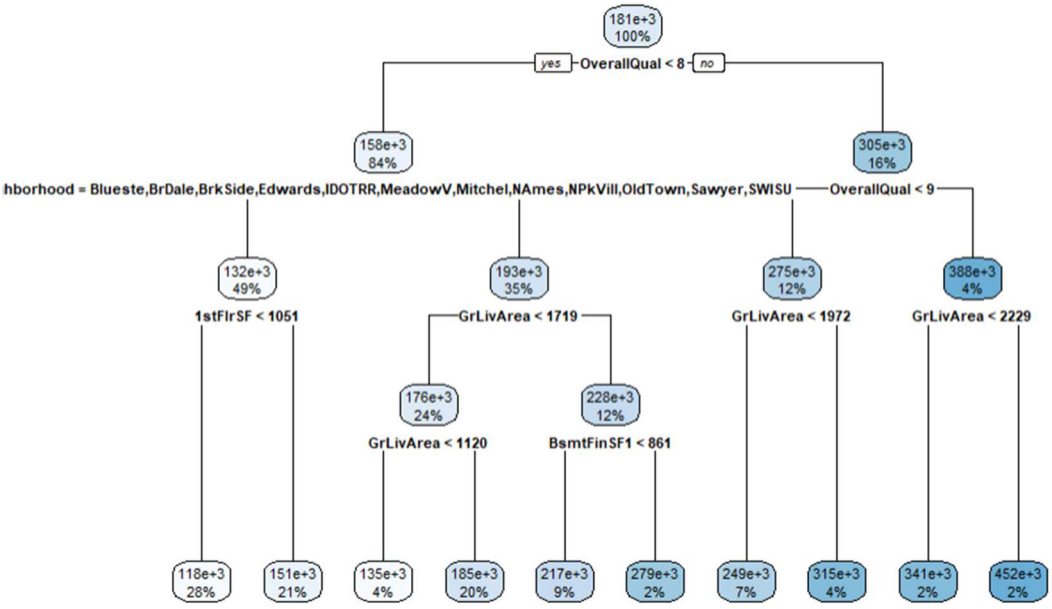
Si OverallQual = 8, entonces GrLivArea (< 1972 vs. ≥ 1972) sigue siendo una variable clave. Casas con más área tienden a costar más (~315,000 USD).

Si OverallQual ≥ 9, la casa es más cara (~388,000 a 450,000 USD), y nuevamente el tamaño (GrLivArea) es crucial.

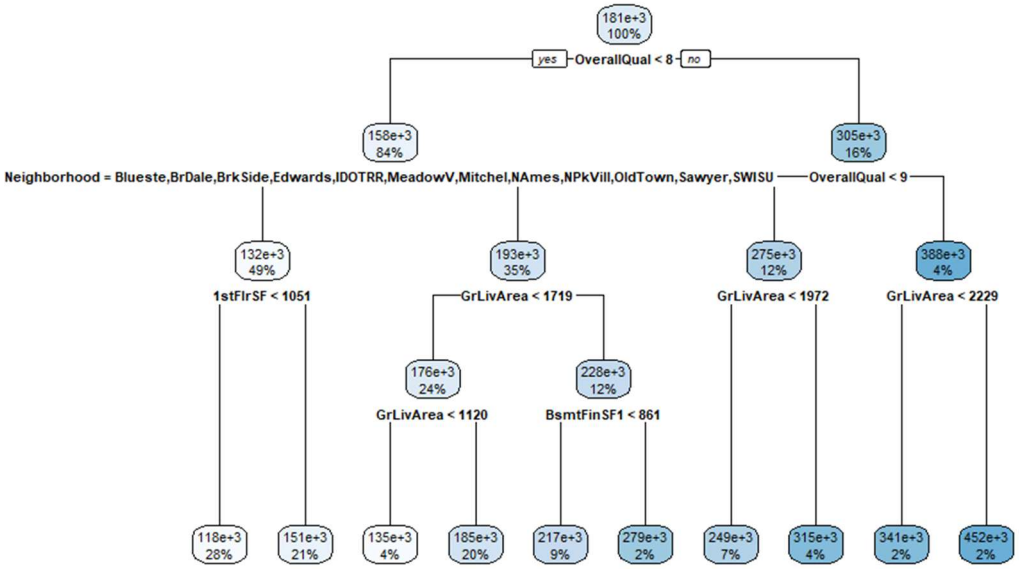
Cambiando profundidades

Vamos a ir cambiando la profundidad del gráfico entre más vamos

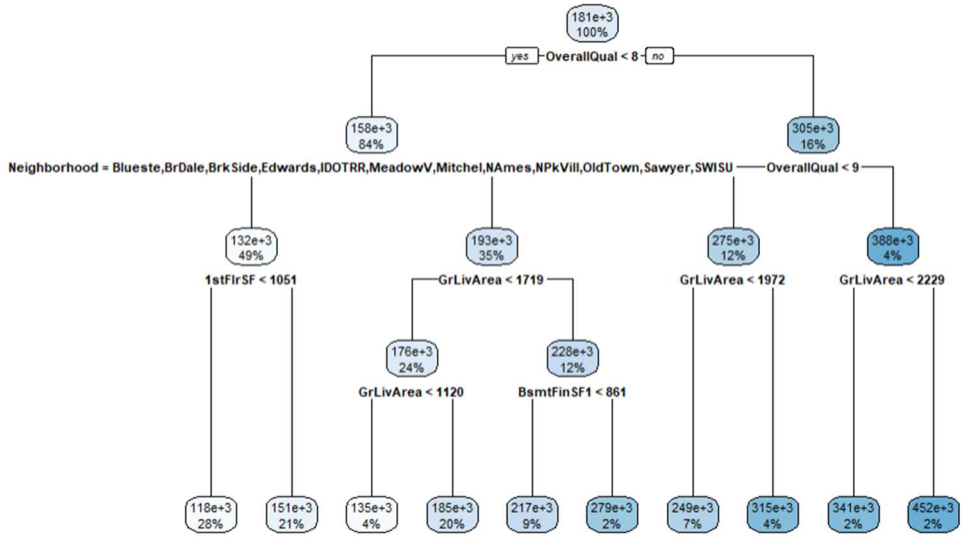
vamos a empezar con depth 3



Ahora sigamos con deapth 4



Ahora sigamos con deapth 5



Lamentablemente no cambia mucho o nada, pero podemos interpretar cosas con lo que hemos aprendido

Analizando los árboles

La calidad general (OverallQual) es el factor más importante

El nodo raíz divide los datos en dos ramas principales basadas en si OverallQual < 8 o no.

Esto indica que la calidad general de la vivienda es un predictor clave del precio de venta.

Dentro de casas con menor calidad (OverallQual < 8), el vecindario juega un rol importante

Si la casa está en ciertos vecindarios (Blueste, BrDale, etc.), el precio tiende a ser más bajo.

Para estos vecindarios, el tamaño de la primera planta (1stFlrSF) afecta el precio.

En casas con mejor calidad (OverallQual ≥ 8), el tamaño de la vivienda es clave

La variable GrLivArea (área habitable sobre el suelo) es un factor determinante en el precio.

Si GrLivArea es mayor a 2229, los precios son significativamente más altos.

Los precios más altos ocurren en casas con alta calidad y gran área habitable

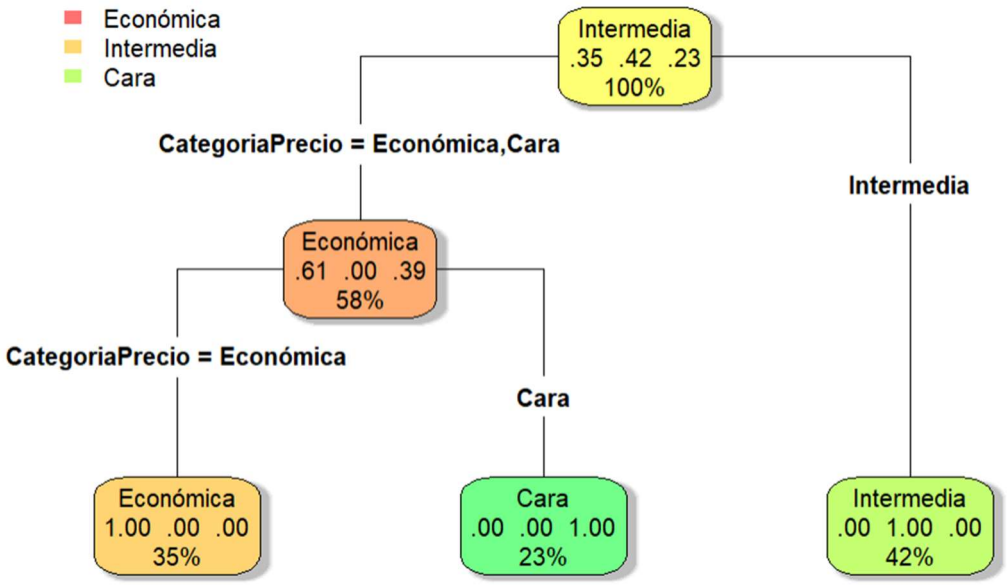
Las casas con OverallQual ≥ 9 y GrLivArea ≥ 2229 tienen los precios más elevados .

Creando varaibles

Vamos a crear variables para que se pueda analizar mejor las cosas

Económica	Intermedia	Cara
518	607	335

ok, ya tenemos los datos, ahora apliquemos a ver que sucede]



Variables como 1stFt SF (tamaño del primer piso), GarageArea (área del garaje), GrLivArea (área habitable) y TotalBsmtSF (área del sótano) determinan la categoría asignada.

Ejemplo: Si 1stFt SF < 1051, el 50% se clasifica como "Económica".

Los porcentajes indican la proporción de propiedades que cumplen ciertas condiciones y caen en una categoría.

Ejemplo: Bajo GarageArea < 301, el 21% se clasifica como "Intermedia".

Con esto en mente logramos apreciar cómo se logran apreciar las nuevas subcategorías creadas gracias a la nueva variable

Haciendo un Random Forest

```
Overall Statistics

               Accuracy : 0.8411
                 95% CI : (0.7728, 0.8954)
    No Information Rate : 0.4901
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.73

 Mcnemar's Test P-Value : 0.2102

Statistics by Class:

                     Class: Económica Class: Intermedia Class: Cara
Sensitivity                   0.57895            0.8378      0.9310
Specificity                   0.97727            0.8571      0.8925
Pos Pred Value                0.78571            0.8493      0.8438
Neg Pred Value                0.94161            0.8462      0.9540
Prevalence                    0.12583            0.4901      0.3841
Detection Rate                0.07285            0.4106      0.3576
Detection Prevalence          0.09272            0.4834      0.4238
Balanced Accuracy             0.77811            0.8475      0.9118
rf variable importance

  only 20 most important variables shown (out of 249)
```

Esto significa que tu modelo clasifica correctamente el 84.11% de las observaciones en el conjunto de prueba. Es una buena precisión, lo que sugiere que el modelo tiene un rendimiento sólido.

95% CI (Intervalo de Confianza del 95%): (0.7728, 0.8954)

Esto indica que, con un 95% de confianza, la precisión real del modelo se encuentra entre el 77.28% y el 89.54%.

No Information Rate (Tasa de No Información): 0.4901

Esta es la precisión que obtendrías si siempre predijeras la clase más frecuente. En este caso, es del 49.01%.

P-Value [Acc > NIR]: < 2e-16

Este valor P es extremadamente bajo, lo que significa que la precisión de tu modelo es significativamente mejor que la tasa de no información. Esto indica que el modelo está aprendiendo patrones reales en los datos

Kappa: 0.73

Mcnemar's Test P-Value: 0.2102

Resumiendo lo aprendido

A lo largo de este proceso logramos deducir muchas cosas, entre una de ellas es que una de las mayores Si que existen correlaciones no vistas anteriormente, por ejemplo en un árbol se logro apreciar como los pisos, áreas de vivienda y áreas extras, lograr hacer que muchas casas fueran categorizadas de maneras muy distintas, al mismo tiempo logramos presenciar como cada árbol evoluciono, Se logró predecir con una incertidumbre bastante moderada, aunque algo alta, cada uno logro hacer una predicción buena y creo que fue un modelado exitoso