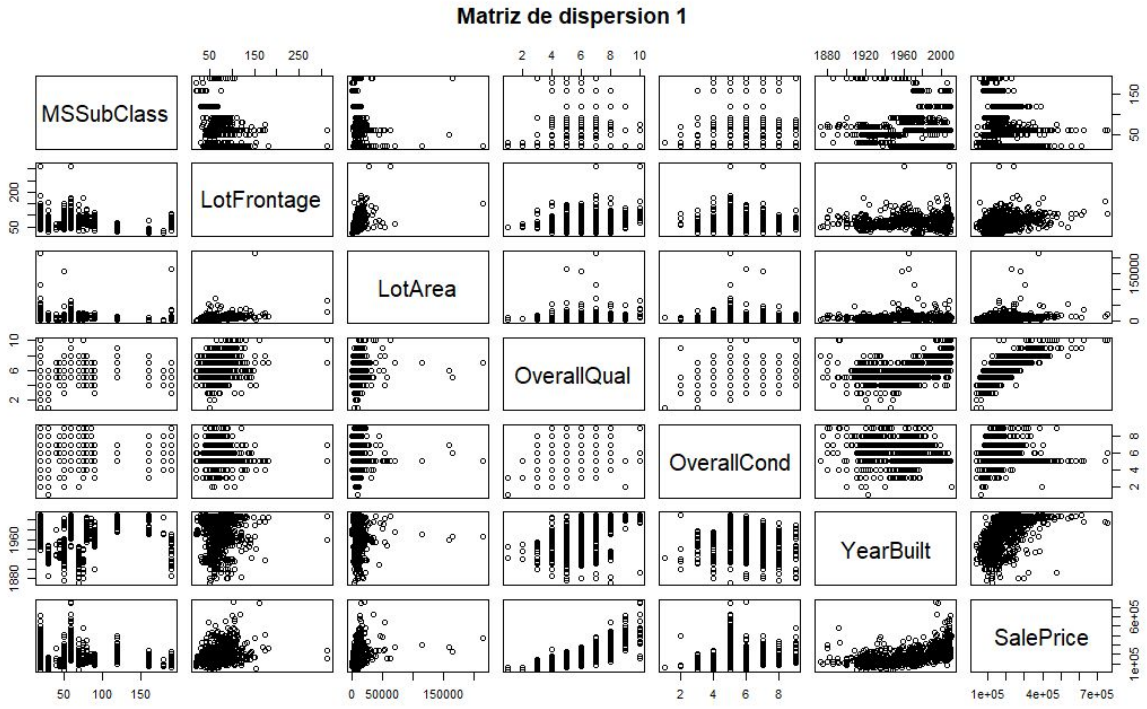
Universidad del Valle de Guatemala
Minería de Datos - Sección 10
María Fernanda Estrada 14198
Christopher Kevin Sandoval 13660
09/03/2020

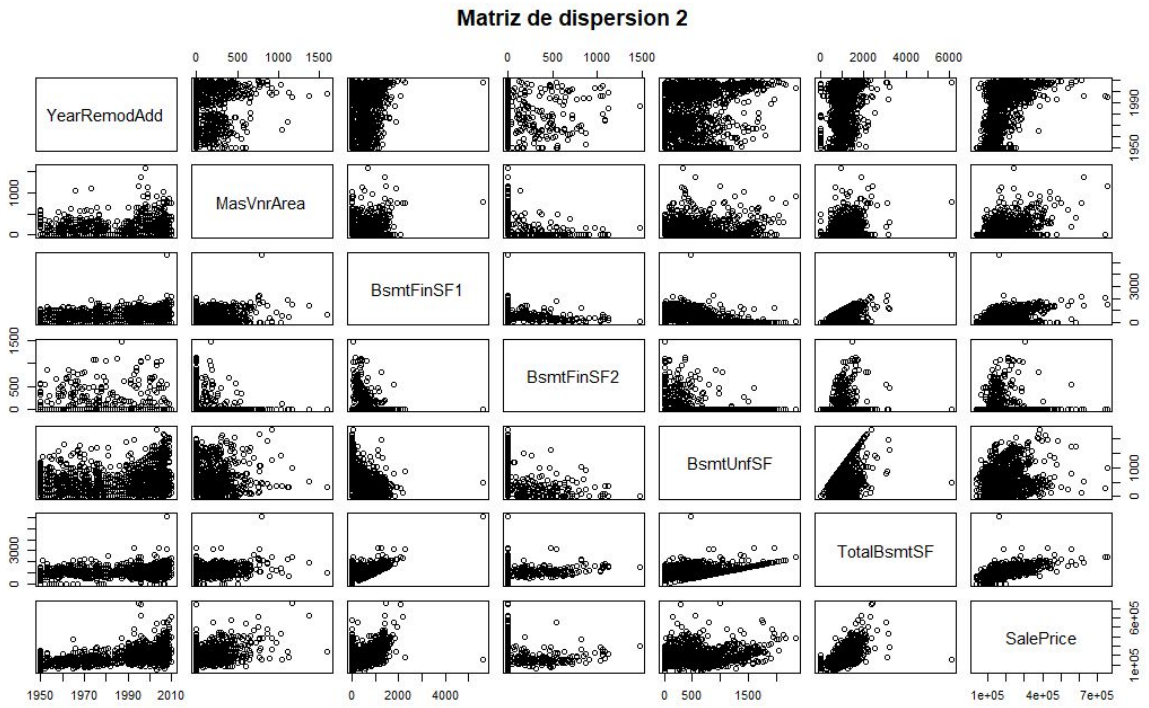
Hoja de Trabajo 3

Análisis exploratorio

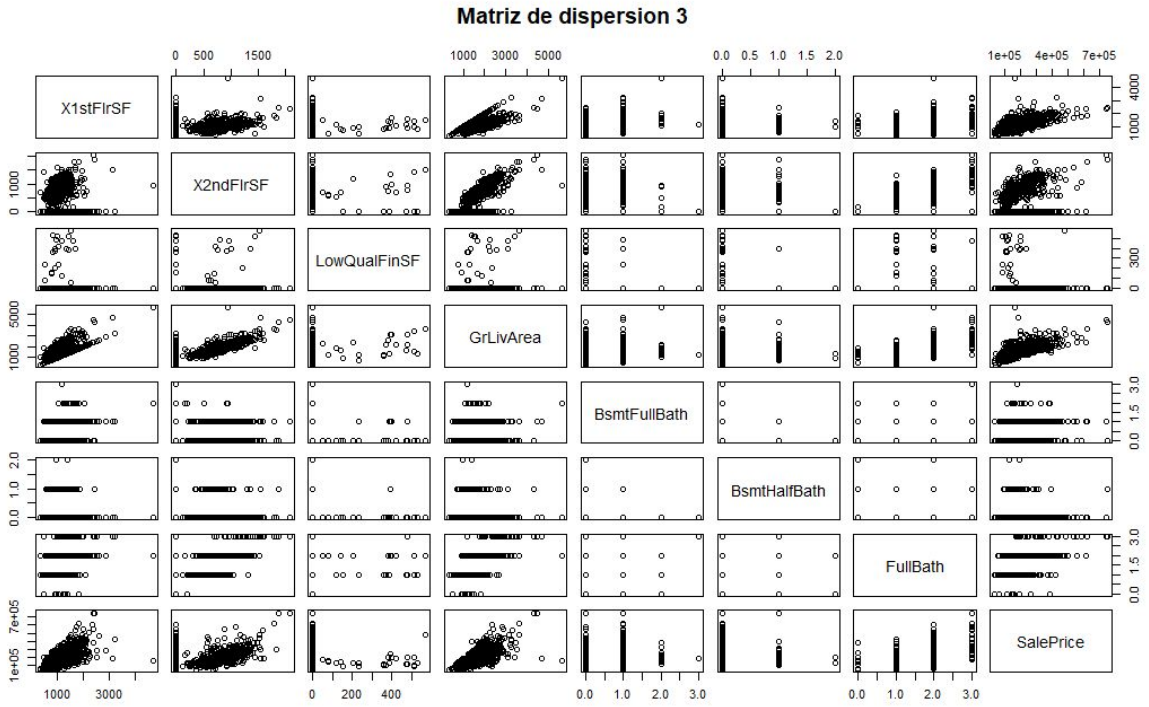
Para poder hacer un análisis con gráficas, el preprocesamiento incluyó descartar las variables no numéricas. Sin embargo, se tuvieron que separar las variables numéricas en tres grupos ya que eran demasiadas para que se entendieran bien en una sola gráfica.
En la primera gráfica, se encontró que sólo dos variables tienen relación con *SalePrice*. Estas variable son: material general y calidad de acabado -*OverallQual*-, y la fecha original de construcción -*YearBuilt*-.



En la segunda gráfica, solamente hay una variable relacionada con *SalePrice*, la cual es el total de pies cuadrados de área de sótano -*TotalBsmtSF*-. Se tomó esta variable, ya que las variables *BsmtFinSFx* tienen directa relación con el total, pero no tanto con *SalePrice* si se consideran individualmente.

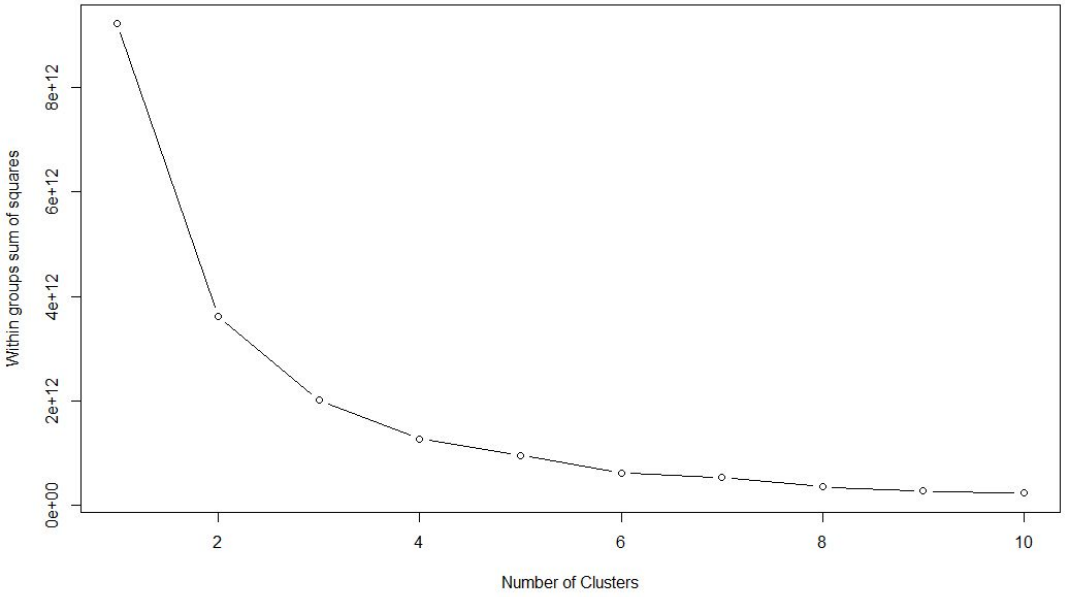


En la tercera gráfica, hay tres variables con relación a *SalePrice*. Estas variables son: los pies cuadrados del primer piso -X1stFlrSF-, los pies cuadrados de superficie habitable por encima del nivel del suelo -*GrLivArea*- y los baños completos -*FullBath*-.

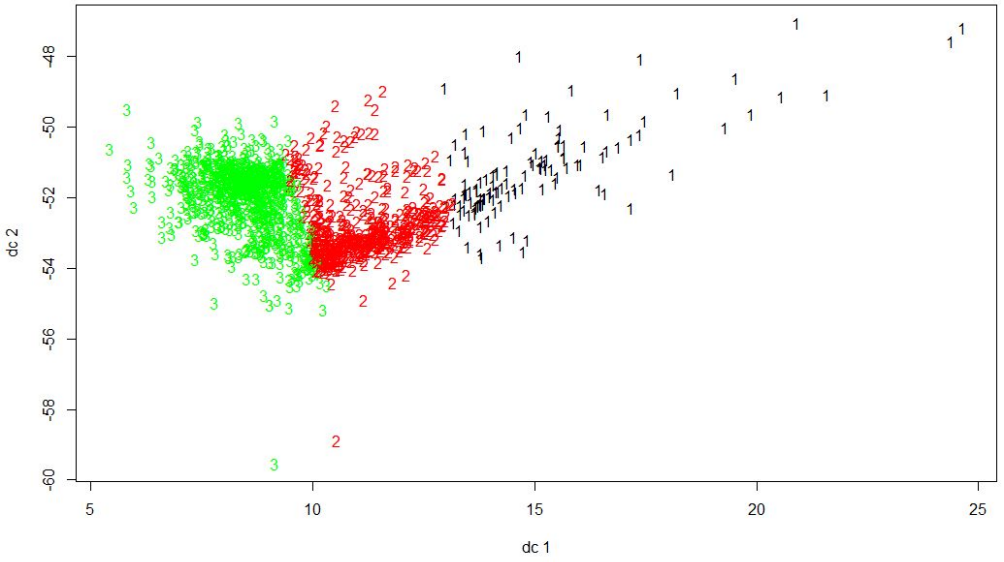


Análisis de los grupos

Para realizar la agrupación, se tomaron únicamente en cuenta las variables antes mencionadas que tienen relación con *SalePrice*. Luego, se realizó una gráfica de codo para saber cuántos grupos podríamos formar. En base a la gráfica, el número ideal de grupos es 3.



A partir de este dato, se utilizó el algoritmo Kmeans para agrupación. El resultado de la silueta de Kmeans fue de 0.56.



A continuación, se presenta un resumen de cada grupo formado, dándo una idea de las características que poseen.
El primer grupo posee las siguientes características:

- Casas caras, siendo el mínimo de $297,000 y el máximo de $755,000.
- La calidad de acabado y material general usado son elevados, siendo casas más finas estéticamente con un promedio de 8.487
- Son casas más recientes, su mediana es el año 2005.
- Son casas más grandes (las que mayor cantidad de pies cuadrados tienen), tomando en cuenta las variables *TotalBsmtSF*, *X1stFlrSF* y *GrLivArea*.
- En promedio, tienen más baños.

```
> summary(cluster1)
   SalePrice       OverallQual      YearBuilt     TotalBsmtSF     X1stFlrSF      GrLivArea
Min.   :297000   Min.   : 7.000   Min.   :1892   Min.   : 728   Min.   :1026   Min.   :1419
1st Qu.:318061   1st Qu.: 8.000   1st Qu.:1996   1st Qu.:1393   1st Qu.:1462   1st Qu.:1902
Median :345000   Median : 8.000   Median :2005   Median :1698   Median :1718   Median :2234
Mean   :374621   Mean   : 8.487   Mean   :1997   Mean   :1668   Mean   :1707   Mean   :2315
3rd Qu.:395000   3rd Qu.: 9.000   3rd Qu.:2007   3rd Qu.:1926   3rd Qu.:1944   3rd Qu.:2622
Max.   :755000   Max.   :10.000   Max.   :2010   Max.   :3200   Max.   :3228   Max.   :4476
    FullBath       cluster
Min.   :0.000   Min.   :1
1st Qu.:2.000   1st Qu.:1
Median :2.000   Median :1
Mean   :2.085   Mean   :1
3rd Qu.:2.000   3rd Qu.:1
Max.   :3.000   Max.   :1
```

El segundo grupo posee las siguientes características:

- Casas intermedias, siendo el mínimo de $173,500 y el máximo de $295,493.
- La calidad de acabado y material general usado son medios, siendo casas semi-finas estéticamente con un promedio de 6.86
- No son casas tan recientes, pero no son tan viejas; su mediana es el año 1999.
- Son casas medianas, tomando en cuenta las variables *TotalBsmtSF*, *X1stFlrSF* y *GrLivArea*.
- En promedio, tienen 1.911 baños; menos que el grupo 1 y más que el grupo 3.

```
> summary(cluster2)
   SalePrice       OverallQual      YearBuilt     TotalBsmtSF     X1stFlrSF      GrLivArea
Min.   :173500   Min.   : 4.00   Min.   :1880   Min.   :   0   Min.   : 495.0   Min.   : 988
1st Qu.:187500   1st Qu.: 6.00   1st Qu.:1977   1st Qu.: 880   1st Qu.: 989.5   1st Qu.:1492
Median :210000   Median : 7.00   Median :1999   Median :1201   Median :1263.5   Median :1668
Mean   :216823   Mean   : 6.86   Mean   :1989   Mean   :1197   Mean   :1276.2   Mean   :1751
3rd Qu.:240000   3rd Qu.: 7.00   3rd Qu.:2005   3rd Qu.:1474   3rd Qu.:1526.2   3rd Qu.:1951
Max.   :295493   Max.   :10.00   Max.   :2009   Max.   :3206   Max.   :3138.0   Max.   :4676
    FullBath       cluster
Min.   :0.000   Min.   :2
1st Qu.:2.000   1st Qu.:2
Median :2.000   Median :2
Mean   :1.911   Mean   :2
3rd Qu.:2.000   3rd Qu.:2
Max.   :3.000   Max.   :2
```

El tercer grupo posee las siguientes características:

- Casas económicas, siendo el mínimo de $34,900 y el máximo de $173,000.
- La calidad de acabado y material general usado son bajos, siendo casas sencillas con un promedio de 5.264
- Son las casas más viejas; su mediana es el año 1959.
- Son casas pequeñas, tomando en cuenta las variables *TotalBsmtSF*, *X1stFlrSF* y *GrLivArea*.

- En promedio, tienen 1.266 baños; siendo el grupo con menos baños completos.

```
> summary(cluster3)
   SalePrice       OverallQual       YearBuilt     TotalBsmtSF       X1stFlrSF
 Min.   : 34900   Min.   : 1.000   Min.   :1872   Min.   :   0.0   Min.   : 334.0
 1st Qu.:113000   1st Qu.: 5.000   1st Qu.:1940   1st Qu.: 715.0   1st Qu.: 832.5
 Median :133700   Median : 5.000   Median :1959   Median : 882.0   Median : 970.0
 Mean   :129855   Mean   : 5.264   Mean   :1956   Mean   : 879.3   Mean   :1010.8
 3rd Qu.:149800   3rd Qu.: 6.000   3rd Qu.:1972   3rd Qu.:1056.0   3rd Qu.:1144.0
 Max.   :173000   Max.   :10.000   Max.   :2009   Max.   :6110.0   Max.   :4692.0
   GrLivArea       FullBath        cluster
 Min.   : 334   Min.   :0.000   Min.   :3
 1st Qu.: 980   1st Qu.:1.000   1st Qu.:3
 Median :1190   Median :1.000   Median :3
 Mean   :1248   Mean   :1.266   Mean   :3
 3rd Qu.:1456   3rd Qu.:2.000   3rd Qu.:3
 Max.   :5642   Max.   :3.000   Max.   :3
```

Creación de la variable respuesta para árbol de clasificación

Según los análisis anteriores, se seleccionó que la variable respuesta para el árbol de clasificación es *SalePrice*, la cual es la única variable que nos indica realmente si una casa es económica, intermedia o cara.
Para establecer los límites de estos tres grupos, se tomó en cuenta el análisis de grupos en el inciso anterior. Los límites de los grupos son:

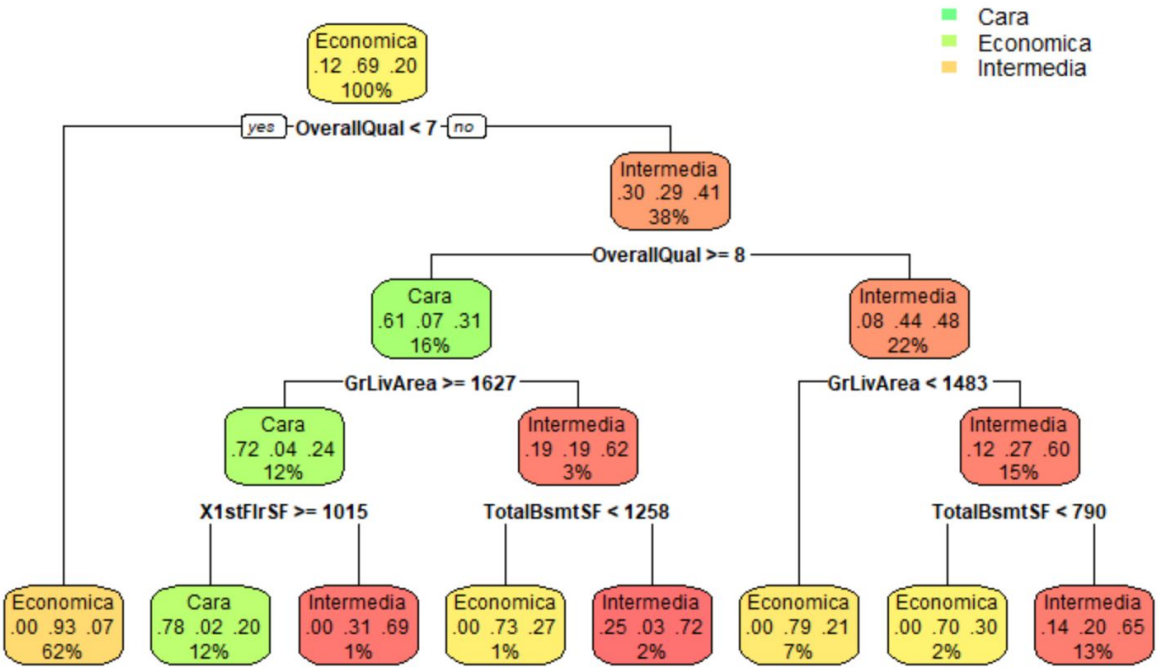
- Económicas: menores a $195,000
- Intermedias: $195,001 a $270,000
- Caras: mayores a $270,001

Conjuntos de entrenamiento y prueba

En teoría, estos grupos se deben crear de forma aleatoria para evitar sesgo de información y que los grupos no representen al resto. En este caso, la página kaggle ya nos proporciona los conjuntos separados. Así que para validar estos grupos, primero se volvieron a unir estas tres tablas. Luego, se separó porcentualmente los datos -60% y 40% (20% para CV) para entrenamiento y prueba, respectivamente- y de manera aleatoria.

Árbol de clasificación

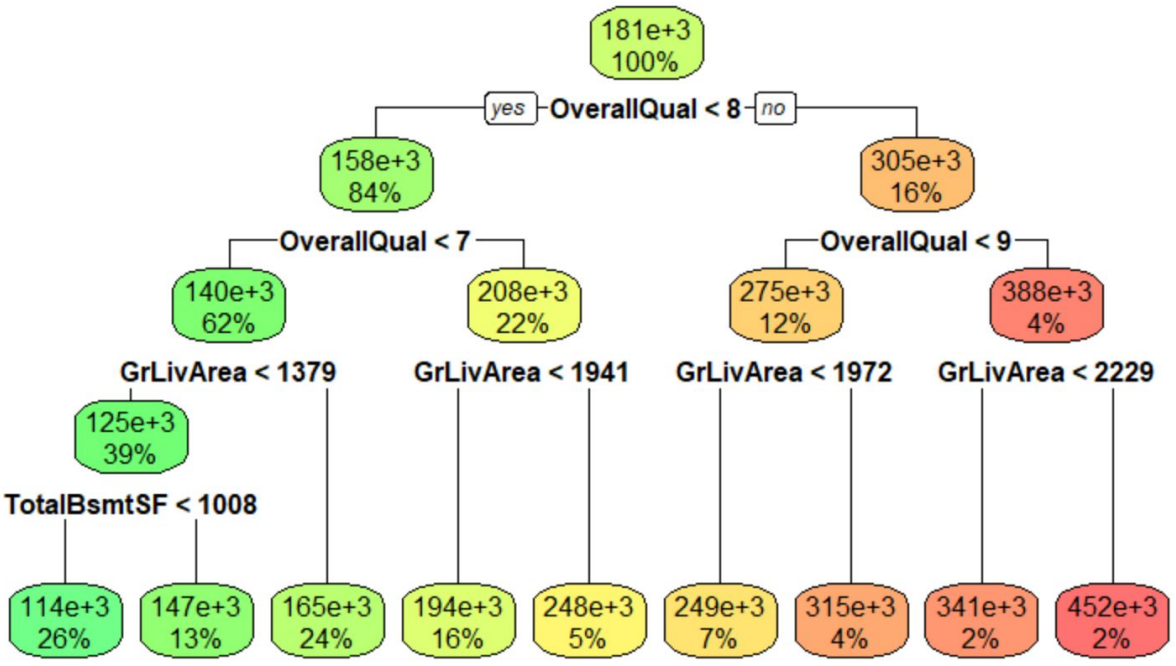
El árbol nos indica que las variables que son importantes a considerar para clasificar una casa son la calidad de materiales/acabados, el tamaño que tiene en ciertas áreas y el año en que fue construida originalmente. La gráfica generada es la siguiente:



Por ejemplo, si la casa tiene calidad mayor o igual que 9, se clasificará como Cara sin considerar otros factores. Al igual que si la calidad es menor a 7 y los pies cuadrados de superficie habitable por encima del nivel del suelo son menores que 1409, se clasificará como Económica.

Árbol de regresión

Este modelo consideró que solamente la calidad y tamaño de ciertas áreas de la casa eran importantes para determinar el precio. Las casas económicas poseen un color más verde amarillento y las más caras ya se colocan en color rojo.
Por ejemplo, las casas con calidad mayor a 9 y con un mayor número de pies cuadrados de superficie habitable por encima del nivel del suelo que 2229 cuestan aproximadamente $452,000. Las casas de calidad menor a 7 y más pequeñas cuestan aproximadamente $114,000.



Análisis de resultados en conjunto de prueba

- Utilización de los modelos con el conjunto de prueba
  - Árbol de clasificación

	OverallQual	TotalBsmtSF	X1stFlrSF	GrLivArea	FullBath	YearBuilt	prediccion
1	5	882	896	896	1	1961	Economica
2	6	1329	1329	1329	1	1958	Economica
3	5	928	928	1629	2	1997	Economica
4	6	926	926	1604	2	1998	Economica
5	8	1280	1280	1280	2	1992	Intermedia
6	6	763	763	1655	2	1993	Economica
7	6	1168	1187	1187	2	1992	Economica
8	6	789	789	1465	2	1998	Economica
9	7	1300	1341	1341	1	1990	Economica
10	4	882	882	882	1	1970	Economica

- Árbol de regresión

	OverallQual	TotalBsmtSF	X1stFlrSF	GrLivArea	FullBath	YearBuilt	prediccion
1	5	882	896	896	1	1961	113919.9
2	6	1329	1329	1329	1	1958	146883.5
3	5	928	928	1629	2	1997	165466.1
4	6	926	926	1604	2	1998	165466.1
5	8	1280	1280	1280	2	1992	249392.5
6	6	763	763	1655	2	1993	165466.1
7	6	1168	1187	1187	2	1992	146883.5
8	6	789	789	1465	2	1998	165466.1
9	7	1300	1341	1341	1	1990	194238.7
10	4	882	882	882	1	1970	113919.9

- Eficiencia árbol de clasificación

```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara          1       115          47
  Economica     2       973         102
  Intermedia    0       180          39

Overall Statistics

               Accuracy : 0.6943
                 95% CI : (0.67, 0.7179)
    No Information Rate : 0.8691
    P-Value [Acc > NIR] : 1
```

El algoritmo tuvo más aciertos en clasificar las casas Económicas; se ha equivocado más al clasificar las casas Intermedias. En general, se obtuvo un 69.43% de precisión en la clasificación.

- Desempeño árbol de regresión

Para calcular el desempeño del árbol de regresión, se comparó la predicción que hizo y el valor real del dataset; se obtuvo el porcentaje de error de cada valor y por último el porcentaje de error promedio de todas las predicciones. El porcentaje de error promedio general fue de 28.61%.

	OverallQual	TotalBsmtSF	X1stFlrSF	GrLivArea	FullBath	YearBuilt	prediccion	real	error
1	5	882	896	896	1	1961	113919.9	169277.1	32.7020905
2	6	1329	1329	1329	1	1958	146883.5	187758.4	21.7699423
3	5	928	928	1629	2	1997	165466.1	183583.7	9.8688479
4	6	926	926	1604	2	1998	165466.1	179317.5	7.7245055
5	8	1280	1280	1280	2	1992	249392.5	150730.1	65.4563350
6	6	763	763	1655	2	1993	165466.1	177151.0	6.5960118
7	6	1168	1187	1187	2	1992	146883.5	172070.7	14.6376839
8	6	789	789	1465	2	1998	165466.1	175111.0	5.5078606
9	7	1300	1341	1341	1	1990	194238.7	162011.7	19.8918017
10	4	882	882	882	1	1970	113919.9	160726.2	29.1217713

```
> error_total
[1] 28.61673
```

Random forest

Al correr el algoritmo de Random Forest para comparar con el modelo de clasificación antes mencionado, se encontró que los resultados fueron bastante similares, con una diferencia del 0.0119% en precisión. Similar al caso anterior, logró clasificar mejor las casas Económicas y peor a las casas Intermedias.

	OverallQual	TotalBsmtSF	X1stFlrSF	GrLivArea	FullBath	YearBuilt	predRF
1	5	882	896	896	1	1961	Economica
2	6	1329	1329	1329	1	1958	Economica
3	5	928	928	1629	2	1997	Economica
4	6	926	926	1604	2	1998	Economica
5	8	1280	1280	1280	2	1992	Economica
6	6	763	763	1655	2	1993	Economica
7	6	1168	1187	1187	2	1992	Economica
8	6	789	789	1465	2	1998	Economica
9	7	1300	1341	1341	1	1990	Economica
10	4	882	882	882	1	1970	Economica

```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara          1       106         53
  Economica     1       955         96
  Intermedia    1       206         39

Overall Statistics

               Accuracy : 0.6824
                 95% CI : (0.6579, 0.7063)
    No Information Rate : 0.869
    P-Value [Acc > NIR] : 1
```