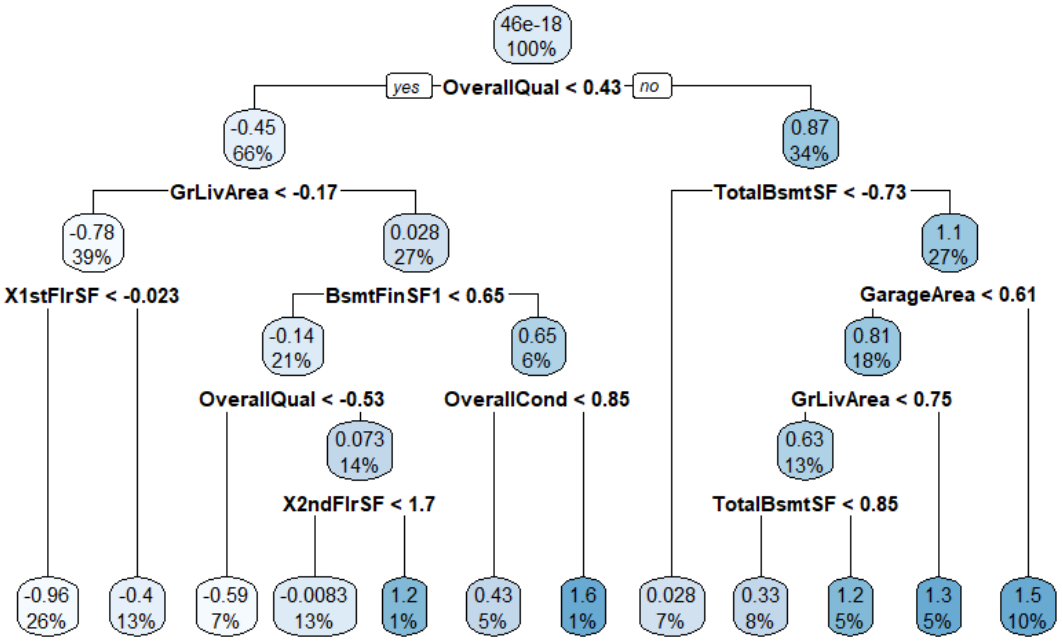
Universidad del Valle de Guatemala
Minería de Datos
Sección 10

Hoja de Trabajo 4

Brian Carrillo - 21108
Diego Alonzo - 20172
Carlos López - 21666

Guatemala, 11 de marzo del 2024

1. Use los mismos conjuntos de entrenamiento y prueba que usó para los modelos de regresión lineal en la hoja de trabajo anterior.
2. Elabore un árbol de regresión para predecir el precio de las casas usando todas las variables.



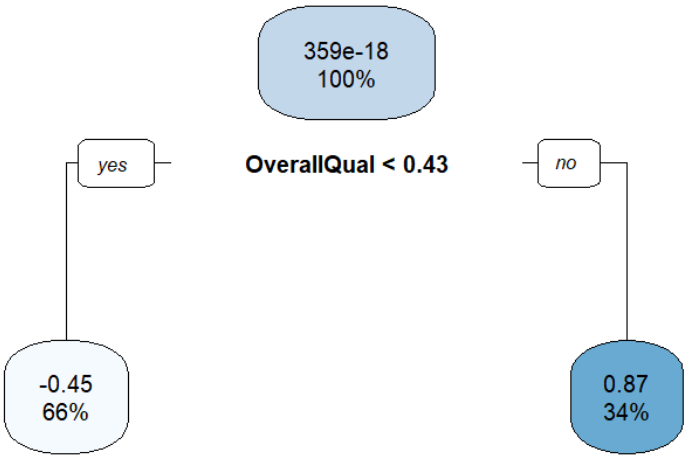
3. Úselo para predecir y analice el resultado. ¿Qué tal lo hizo?

MSE: 0.48
RMSE: 0.69
R^2: 0.52

A partir del coeficiente de determinación obtenido, se puede decir que el árbol de regresión logra explicar solamente el 52% de la varianza de los datos. El MSE y RMSE determinan que el error del modelo es alto, puesto que el conjunto de datos utilizados se encuentra normalizado. Lo idóneo es que estos valores se encuentren lo más cerca de 0 posible. Algunas mejoras podrían ser:

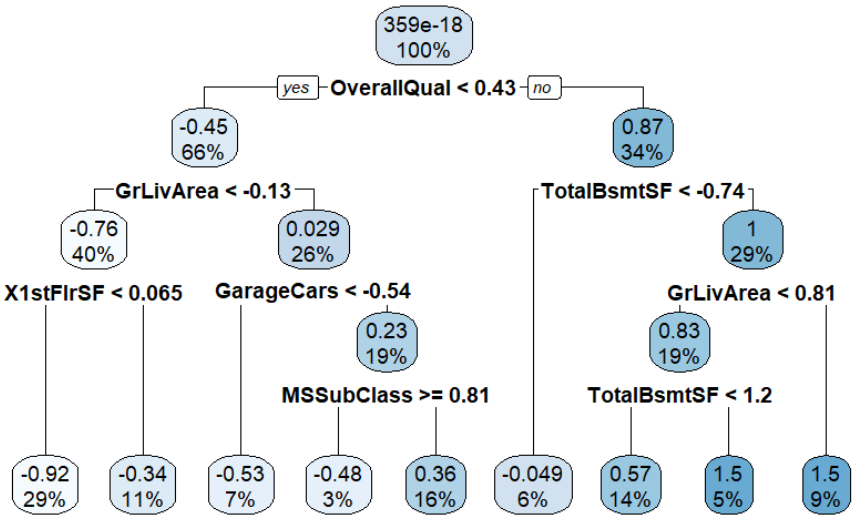
1. Revisar la profundidad del árbol
2. Probar con otros modelos.
3. Validación cruzada.

4. Haga, al menos, 3 modelos más cambiando el parámetro de la profundidad del árbol. ¿Cuál es el mejor modelo para predecir el precio de las casas?
maxdepth = 1



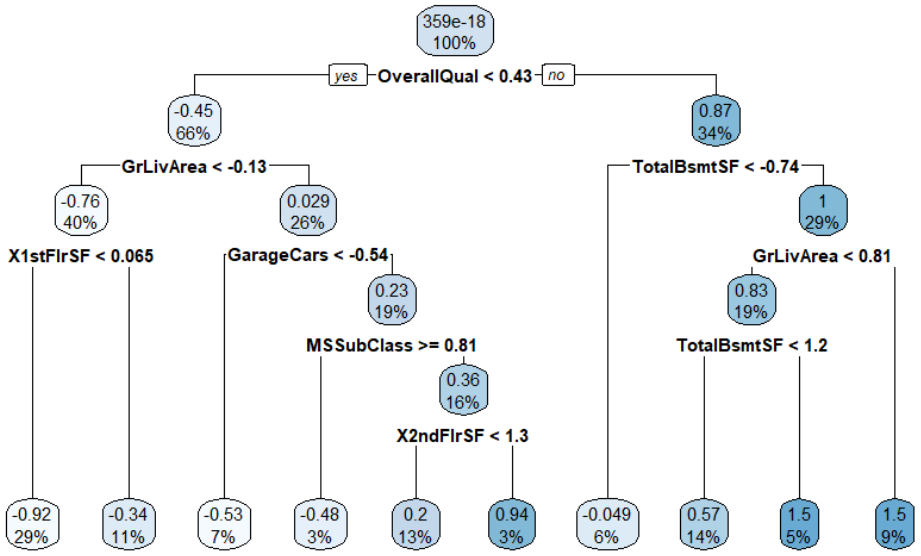
MSE: 0.61
RMSE: 0.78
R^2: 0.39

maxdepth = 2



MSE: 0.37
RMSE: 0.61
R^2: 0.63

maxdepth = 5



MSE: 0.37
RMSE: 0.61
R^2: 0.62

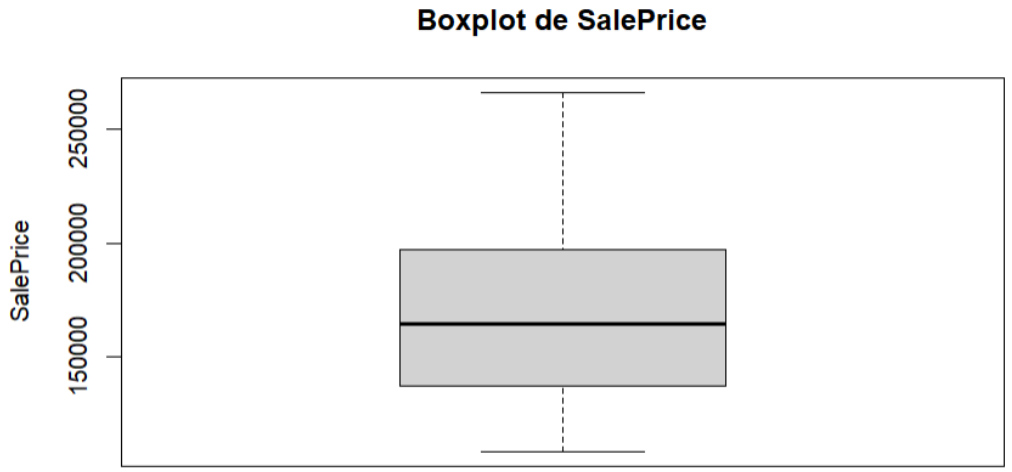
El mejor modelo para predecir el precio de las casas es el que posee un depth de 2, puesto que llega a explicar hasta un 63% de la varianza de los datos.

5. Compare los resultados con el modelo de regresión lineal de la hoja anterior, ¿cuál lo hizo mejor?

El coeficiente de determinación del modelo de regresión lineal univariado de la hoja anterior es de 0.52. Mientras que el coeficiente de determinación del modelo de regresión lineal multivariado es de 0.79. Por lo tanto se puede afirmar que el modelo de regresión lineal multivariado es mejor que el modelo de árbol de regresión. Sin embargo, este último presenta mejores predicciones que las realizadas por el modelo de regresión lineal univariado.

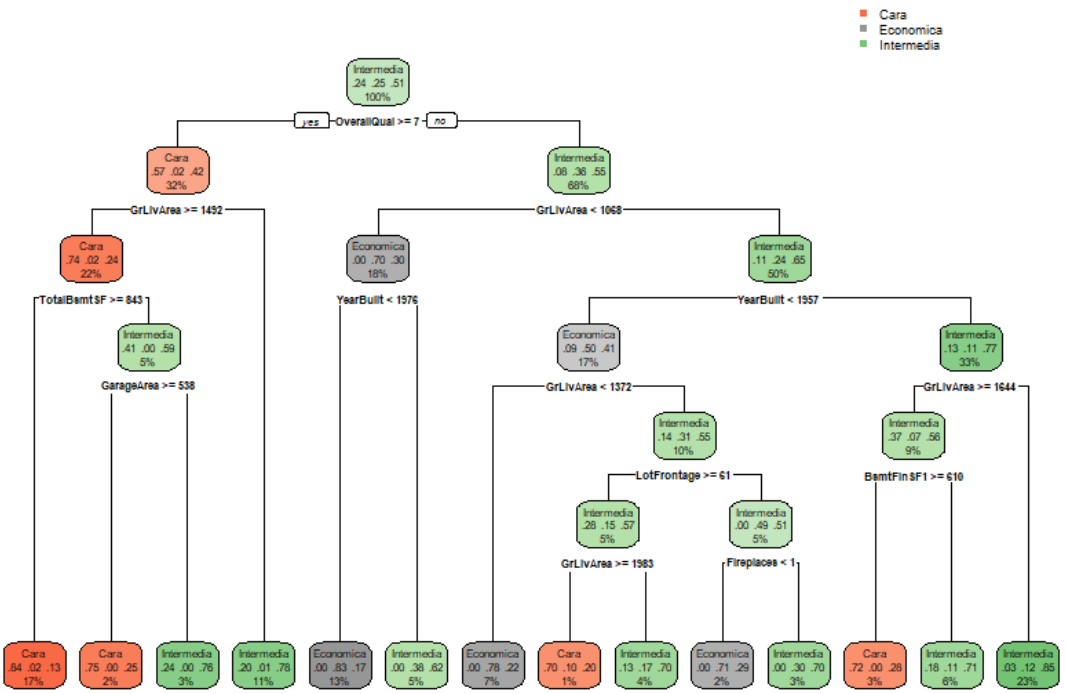
6. Dependiendo del análisis exploratorio elaborado cree una variable respuesta que le permita clasificar las casas en Económicas, Intermedias o Caras. Los límites de estas clases deben tener un fundamento en la distribución de los datos de precios, y estar bien explicados.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
108480  137500  164600  170472  196875  266000
```



La variable de respuesta fue asignada en base al primer y tercer cuartil de la distribución del precio de venta. Esta clasificación permite una distribución equitativa de las casas en cada categoría, reflejando una variedad de opciones para diferentes segmentos. Además, los cuartiles son resistentes a valores atípicos extremos, lo que puede hacer que la clasificación sea más representativa de la mayoría de las casas en el conjunto de datos. Se utilizaron el primer y tercer cuartil como límites. Las casas con SalePrice por debajo del primer cuartil se consideraron "Económicas", las casas por encima del tercer cuartil se consideraron "Caras", y las que están entre ambos son consideradas "Intermedias".

7. Elabore un árbol de clasificación utilizando la variable respuesta que creó en el punto anterior. Explique los resultados a los que llega. Muestre el modelo gráficamente. Recuerde que la nueva variable respuesta es categórica, pero se generó a partir de los precios de las casas, no incluya el precio de venta para entrenar el modelo.



El árbol de decisión obtenido refleja el uso de variables cuantitativas como los predictores en cada nivel. Podemos visualizar que la mayor parte de las hojas pertenecen a predicciones de casas con un precio intermedio, por lo que es posible inducir que gran parte de las predicciones serán de dicha categoría.

8. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar. 9. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión para el árbol de clasificación. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.

```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara         61         4         23
  Economica     0        46         15
  Intermedia   29        31        114

Overall Statistics

               Accuracy : 0.6842
                 95% CI : (0.6304, 0.7346)
    No Information Rate : 0.4706
    P-Value [Acc > NIR] : 7.394e-15

                  Kappa : 0.4933

 Mcnemar's Test P-Value : 0.0165
```

```
Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.6778           0.5679            0.7500
Specificity               0.8841           0.9380            0.6491
Pos Pred Value            0.6932           0.7541            0.6552
Neg Pred Value            0.8766           0.8664            0.7450
Prevalence                0.2786           0.2508            0.4706
Detection Rate            0.1889           0.1424            0.3529
Detection Prevalence      0.2724           0.1889            0.5387
Balanced Accuracy         0.7809           0.7530            0.6996
```

Según los resultados de la matriz de confusión, el árbol tuvo una precisión de 0.68. Tuvo mayor cantidad de errores al predecir una casa como categoría intermedia. Esto concuerda con la observación realizada anteriormente, en la que se veía una clara tendencia en clasificaciones de tipo “intermedia”. En general, el modelo parece tener un rendimiento moderado con una exactitud del 68.42% y un valor de kappa de 0.4933. La sensibilidad es más alta para la clase "Intermedia", lo que indica que el modelo es mejor identificando esta clase correctamente en comparación con las otras. La especificidad es alta para la clase "Economica", lo que sugiere que el modelo es bueno para identificar las no económicas.

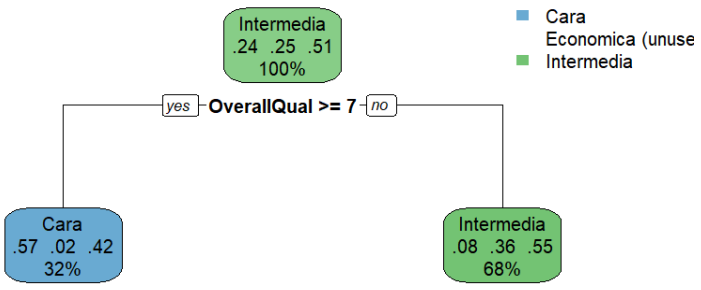
10. Entrene un modelo usando validación cruzada, prediga con él. ¿le fue mejor que al modelo anterior?

```
Confusion Matrix and Statistics

           Reference
Prediction   Cara Economica Intermedia
  Cara         58         2         24
  Economica     0        48         18
  Intermedia   32        31        110

Overall Statistics

               Accuracy : 0.6687
                 95% CI : (0.6145, 0.7199)
    No Information Rate : 0.4706
    P-Value [Acc > NIR] : 5.532e-13

                  Kappa : 0.4693

 Mcnemar's Test P-Value : 0.08611

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.6444           0.5926            0.7237
Specificity               0.8884           0.9256            0.6316
Pos Pred Value            0.6905           0.7273            0.6358
Neg Pred Value            0.8661           0.8716            0.7200
Prevalence                0.2786           0.2508            0.4706
Detection Rate            0.1796           0.1486            0.3406
Detection Prevalence      0.2601           0.2043            0.5356
Balanced Accuracy         0.7664           0.7591            0.6776
```

La precisión y el valor de kappa se mantuvieron similares a los resultados anteriores.

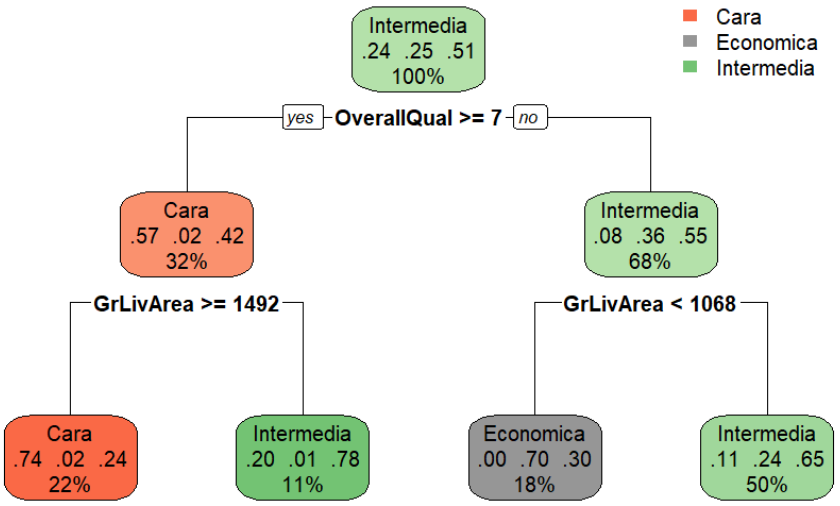
11. Haga al menos, 3 modelos más cambiando la profundidad del árbol. ¿Cuál funcionó mejor?

maxdepth = 1



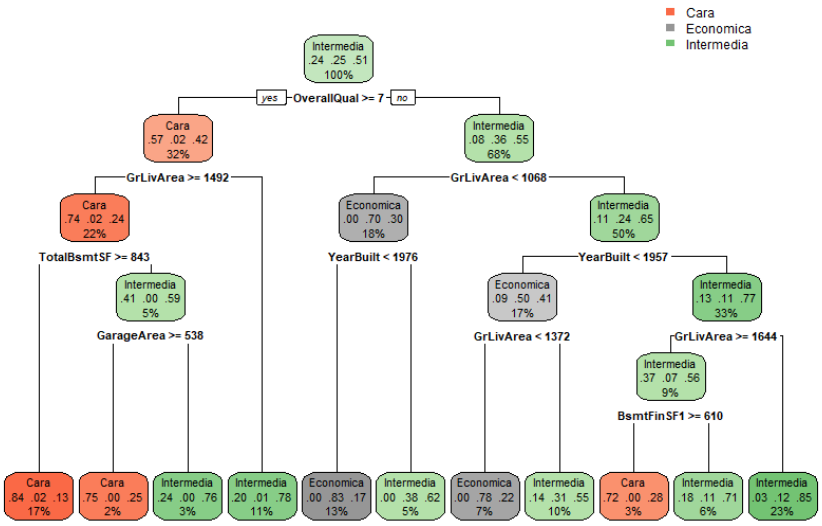
```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara         74         5          49
  Economica     0         0           0
  Intermedia   16        76         103

Overall Statistics

               Accuracy : 0.548
                 95% CI : (0.4919, 0.6032)
    No Information Rate : 0.4706
    P-Value [Acc > NIR] : 0.00318

                  Kappa : 0.2535

 Mcnemar's Test P-Value : < 2e-16

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.8222           0.0000             0.6776
Specificity               0.7682           1.0000             0.4620
Pos Pred Value            0.5781              NaN             0.5282
Neg Pred Value            0.9179           0.7492             0.6172
Prevalence                0.2786           0.2508             0.4706
Detection Rate            0.2291           0.0000             0.3189
Detection Prevalence      0.3963           0.0000             0.6037
Balanced Accuracy         0.7952           0.5000             0.5698
```

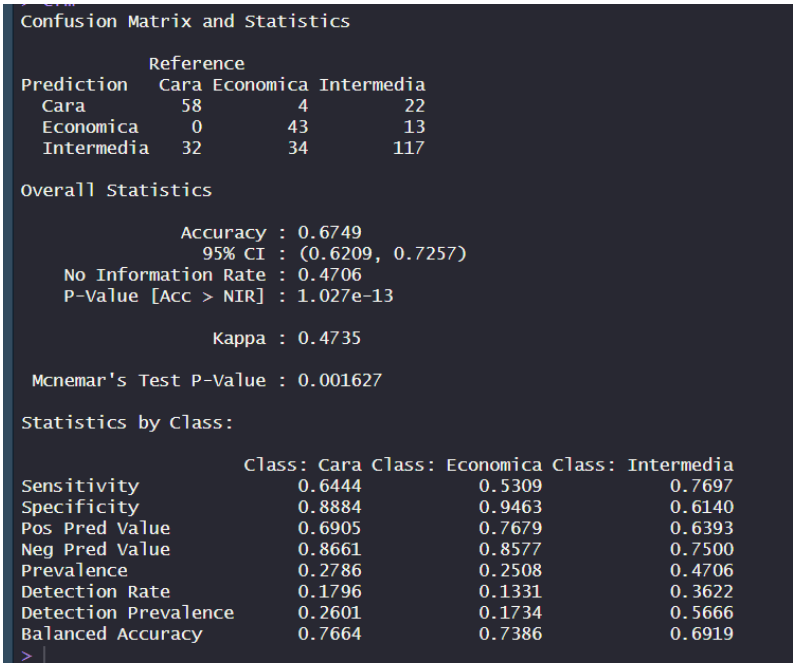
maxdepth = 2



```
Confusion Matrix and Statistics

          Reference
Prediction   Cara Economica Intermedia
  Cara         58         2          24
  Economica     0        35          14
  Intermedia   32        44         114

Overall Statistics

               Accuracy : 0.6409
                 95% CI : (0.5859, 0.6932)
    No Information Rate : 0.4706
    P-Value [Acc > NIR] : 5.536e-10

                  Kappa : 0.4138

 Mcnemar's Test P-Value : 0.0003214

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.6444           0.4321             0.7500
Specificity               0.8884           0.9421             0.5556
Pos Pred Value            0.6905           0.7143             0.6000
Neg Pred Value            0.8661           0.8321             0.7143
Prevalence                0.2786           0.2508             0.4706
Detection Rate            0.1796           0.1084             0.3529
Detection Prevalence      0.2601           0.1517             0.5882
Balanced Accuracy         0.7664           0.6871             0.6528
>
```

maxdepth = 5





```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara         58         4         22
  Economica     0        43         13
  Intermedia   32        34        117

Overall Statistics

               Accuracy : 0.6749
                 95% CI : (0.6209, 0.7257)
    No Information Rate : 0.4706
    P-Value [Acc > NIR] : 1.027e-13

                  Kappa : 0.4735

 Mcnemar's Test P-Value : 0.001627

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.6444           0.5309            0.7697
Specificity               0.8884           0.9463            0.6140
Pos Pred Value            0.6905           0.7679            0.6393
Neg Pred Value            0.8661           0.8577            0.7500
Prevalence                0.2786           0.2508            0.4706
Detection Rate            0.1796           0.1331            0.3622
Detection Prevalence      0.2601           0.1734            0.5666
Balanced Accuracy         0.7664           0.7386            0.6919
>
```

El mejor modelo para predecir el precio de las casas es el que posee un depth de 2, puesto que su precisión fue de 0.67.

12. Repita los análisis usando random forest como algoritmo de predicción, explique sus resultados comparando ambos algoritmos.

```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara         68         0         10
  Economica     0        57         14
  Intermedia   22        24        128

Overall Statistics

               Accuracy : 0.7833
                 95% CI : (0.7343, 0.827)
    No Information Rate : 0.4706
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6527

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.7556           0.7037             0.8421
Specificity               0.9571           0.9421             0.7310
Pos Pred Value            0.8718           0.8028             0.7356
Neg Pred Value            0.9102           0.9048             0.8389
Prevalence                0.2786           0.2508             0.4706
Detection Rate            0.2105           0.1765             0.3963
Detection Prevalence      0.2415           0.2198             0.5387
Balanced Accuracy         0.8563           0.8229             0.7865
```

El resultado de precisión es de 0.78, mayor que todos los modelos realizados anteriormente. El algoritmo Random Forest crea muchos árboles (es decir, un "bosque") y toma el promedio de sus predicciones. Al promediar varios árboles, se reduce la varianza del modelo y se previene el sobreajuste, que es una tendencia común en los árboles de decisión individuales, especialmente si son muy profundos.

Cada árbol en un bosque aleatorio se entrena en una muestra aleatoria de los datos (con reemplazo), y sólo se considera un subconjunto aleatorio de las variables en cada división. Esto asegura que los árboles sean diferentes entre sí, lo que hace que el modelo sea más robusto a los cambios y capaz de generalizar mejor a nuevos datos. Aunque los árboles individuales pueden tener un alto sesgo si son muy simples, o alta varianza si son muy complejos, el bosque aleatorio puede equilibrar esto al combinar los resultados de muchos árboles con diferentes niveles de sesgo y varianza.