Universidad del Valle de Guatemala
Minería de Datos - Sección 10
María Fernanda Estrada 14198
Christopher Kevin Sandoval 13660
27/03/2020

Hoja de Trabajo 5

Análisis del modelo

1. Naive Bayes

   Se utilizó el mismo método de clasificación de la hoja de trabajo 3 para determinar si la casa era Económica, Intermedia o Cara; es decir, la misma variable respuesta con los mismos límites. Luego, para aplicar Naive Bayes se utilizaron las mismas variables que el modelo de regresión lineal, las cuales eran: MSSubClasss, OverallCond, YearBuilt, BsmtFinSF1, X2ndFlrSF, BsmtFullBath, BedroomAbvGr y SceenPorch. Según la matriz de confusión presentada más adelante, el modelo no hace overfitting.

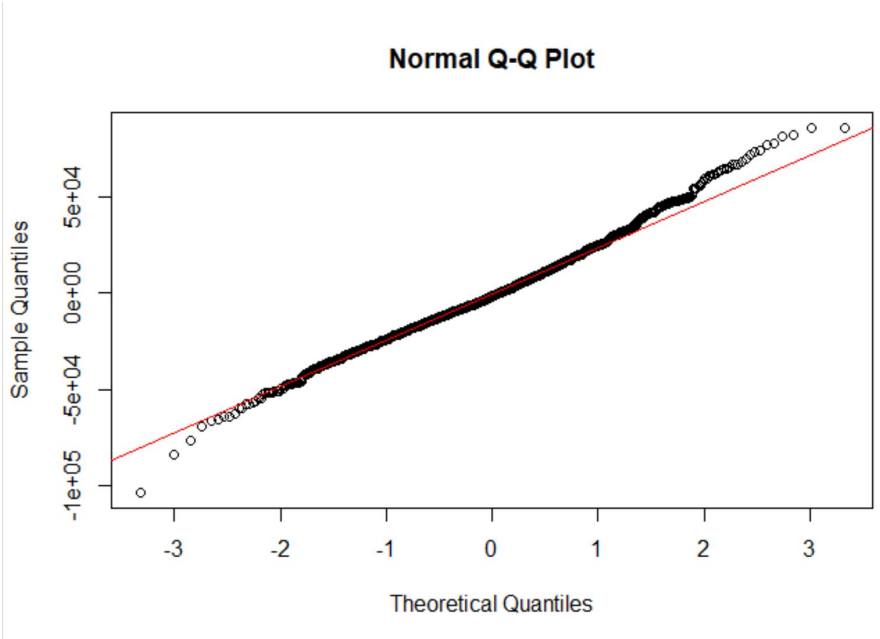
```
> modelo

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
      Cara  Economica Intermedia
 0.1164384  0.6869863  0.1965753

Conditional probabilities:
            MSSubClass
Y                [,1]     [,2]
  Cara       48.50000 28.67823
  Economica  59.41675 46.23968
  Intermedia 53.06620 32.74727

            OverallCond
Y                [,1]      [,2]
  Cara       5.335294 0.9847603
  Economica  5.664008 1.1827818
  Intermedia 5.407666 0.8676452

            YearBuilt
Y                [,1]     [,2]
  Cara       1996.153 20.45344
  Economica  1961.876 28.34839
  Intermedia 1989.348 24.19481

            BsmtFinSF1
Y                [,1]     [,2]
  Cara       851.6000 590.4063
  Economica  363.8953 388.3505
  Intermedia 480.6794 451.9425

            X2ndFlrSF
Y                [,1]     [,2]
  Cara       568.3059 604.3718
  Economica  262.8375 351.7875
  Intermedia 510.0035 492.0160
```

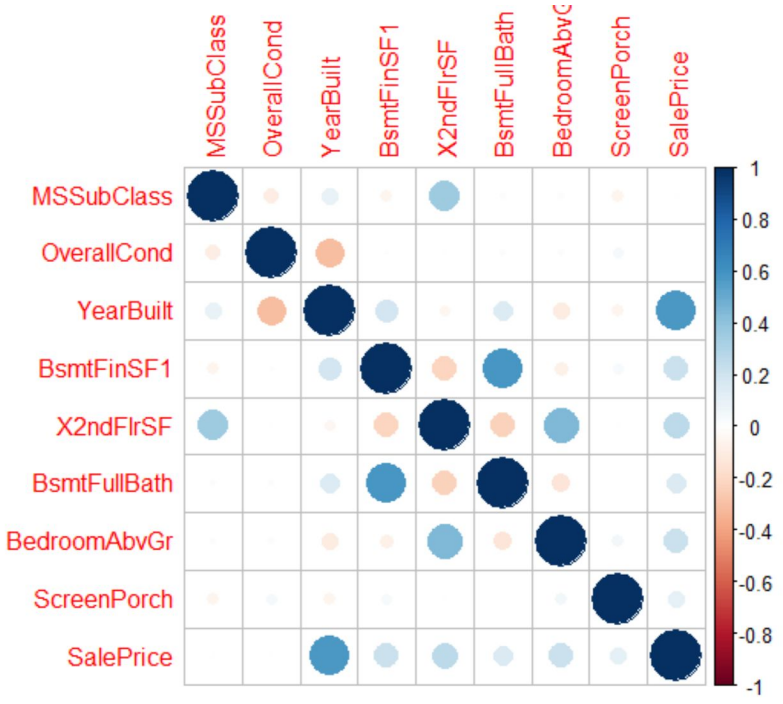
2. Validación cruzada
   El modelo de validación cruzada necesita, aparte de los datos de prueba mencionados en Naive Bayes, el conjunto total de los datos. Es por esto, que se juntaron los datos de test y de training para ello. También, se indicó que el número de folds fuera de 10. Según la matriz de confusión presentada más adelante, este modelo predice mejor que Naive Bayes.

```
> modeloCaret
Naive Bayes

2917 samples
   8 predictor
   3 classes: 'Cara', 'Economica', 'Intermedia'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2626, 2624, 2626, 2626, 2625, 2625, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.7836947  0.3486760
   TRUE      0.8070072  0.2850543
```

Análisis de las variables

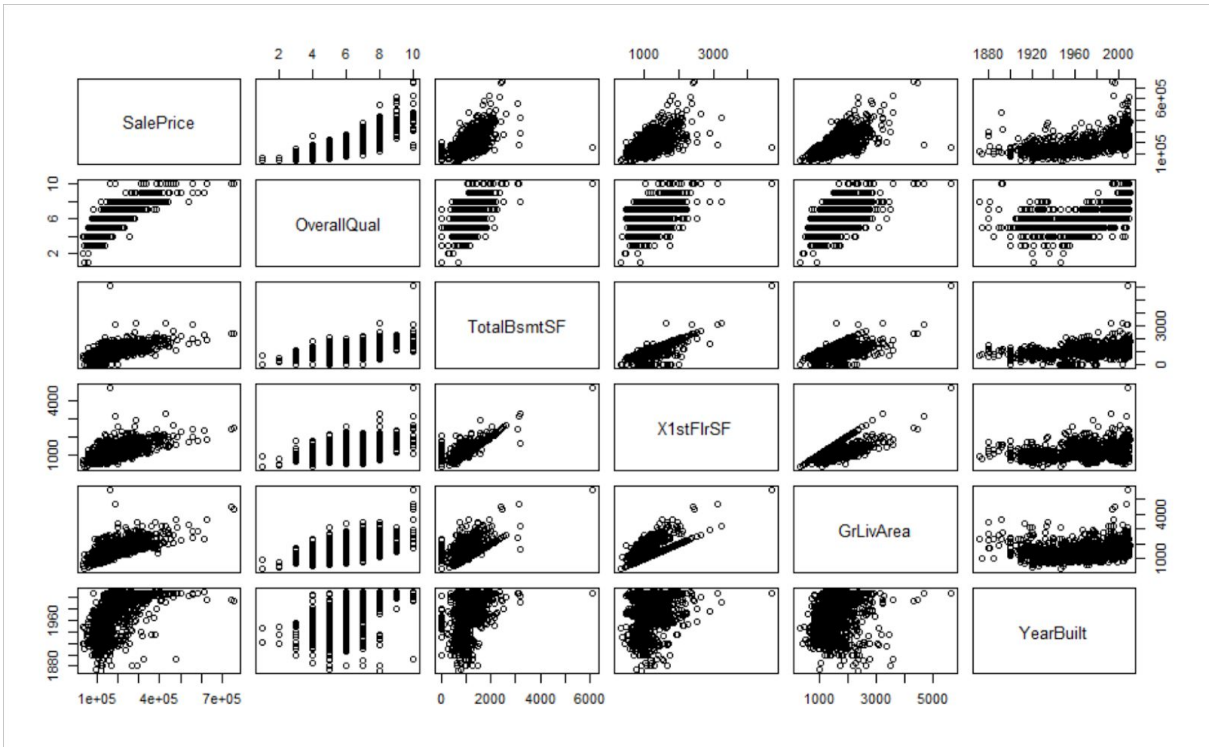
Las pruebas de normalidad sobre los datos produjeron las siguientes gráficas. Primero, se observa que las variables son normales.





Observamos que las relaciones no presentan relación entre ellas, indicando que no afectará el modelo.

Ahora revisamos que no exista multicolinealidad entre las variables. Se observa en la gráfica siguiente que todas presentan relación lineal y que todas aportan al modelo.



Aplicación del modelo al conjunto de prueba

1. Naive Bayes
   Al aplicar el modelo de Naive Bayes en el conjunto de prueba, se dieron los siguientes resultados. El modelo clasificó la mayoría de casas como Económicas.

```
> summary(predBayes)
     Cara  Economica Intermedia
      181       1052        226
```

2. Validación cruzada
   Al aplicar la validación cruzada en el conjunto indicado, se dieron los siguientes resultados. Al igual que en Naive Bayes, la mayoría de casas fueron clasificadas como Económicas.

```
> summary(prediccionCaret)
     Cara  Economica Intermedia
        4       1353        100
```

Matriz de confusión

1. Naive Bayes
   Para determinar la eficiencia de Naive Bayes, se utilizó una matriz de confusión que compara cuánto se equivocó de los datos reales. Se obtuvo un accuracy del **69.9%**, indicando que el modelo predice bastante bien la clasificación de las casas y no hace overfitting.
   Por otro lado, se puede observar que el algoritmo se equivocó más en decir que una casa era Intermedia cuando realmente era Económica. Tuvo muchos aciertos en clasificar una casa como Económica, siendo ésta Económica. En este caso, el error más importante o el que más afectaría sería clasificar una casa como Económica cuando realmente es Cara.

```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara          1       143         37
  Economica     2       959         91
  Intermedia    0       166         60

Overall Statistics

               Accuracy : 0.6991
                 95% CI : (0.6748, 0.7226)
    No Information Rate : 0.8691
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1479
```

2. Validación cruzada
   En este modelo se obtuvo un accuracy del **90.1%**, lo cual indica que predice mucho mejor que Naive Bayes e igualmente sin tener overfitting. Este modelo se equivocó más en clasificar una casa como Económica cuando realmente era Intermedia, un error que no afectaría mucho en la vida real. Al igual que en Naive Bayes, tuvo más aciertos en clasificar casas Económicas.

```
> confusionMatrix(prediccionCaret,data_test_filtered$Class)
Confusion Matrix and Statistics

             Reference
Prediction    Cara Economica Intermedia
  Cara           1         1           2
  Economica      1      1239         113
  Intermedia     0        27          73

Overall Statistics

               Accuracy : 0.9012
                 95% CI : (0.8847, 0.916)
    No Information Rate : 0.8696
    P-Value [Acc > NIR] : 0.000124

                  Kappa : 0.4617
```

Comparación de modelos

Si se compara el modelo de Naive Bayes con el del árbol de clasificación, se puede concluir que son muy similares en cuanto a resultados y desempeño; la diferencia de accuracy en la matriz de confusión fue de 0.47%. Sin embargo, el modelo de validación cruzada superó en un 20.69% al del árbol de clasificación. Además de tener un accuracy mejor, las líneas de código son menos y por ende se tarda menos en procesar.