Universidad del Valle de Guatemala
Minería de Datos - Sección 10
María Fernanda Estrada 14198
Christopher Kevin Sandoval 13660
16/04/2020

Hoja de Trabajo 6

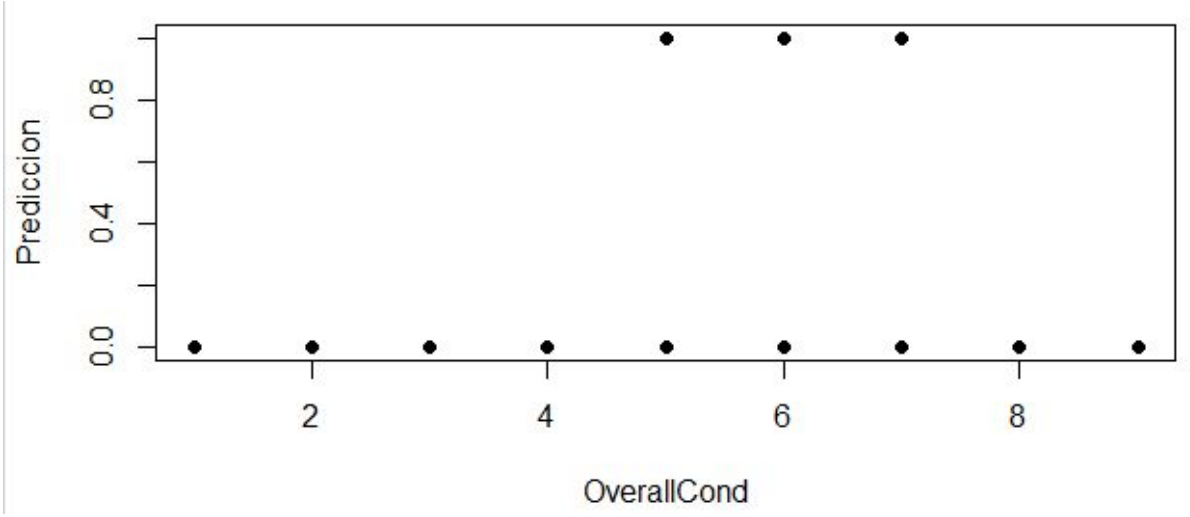
Análisis del modelo

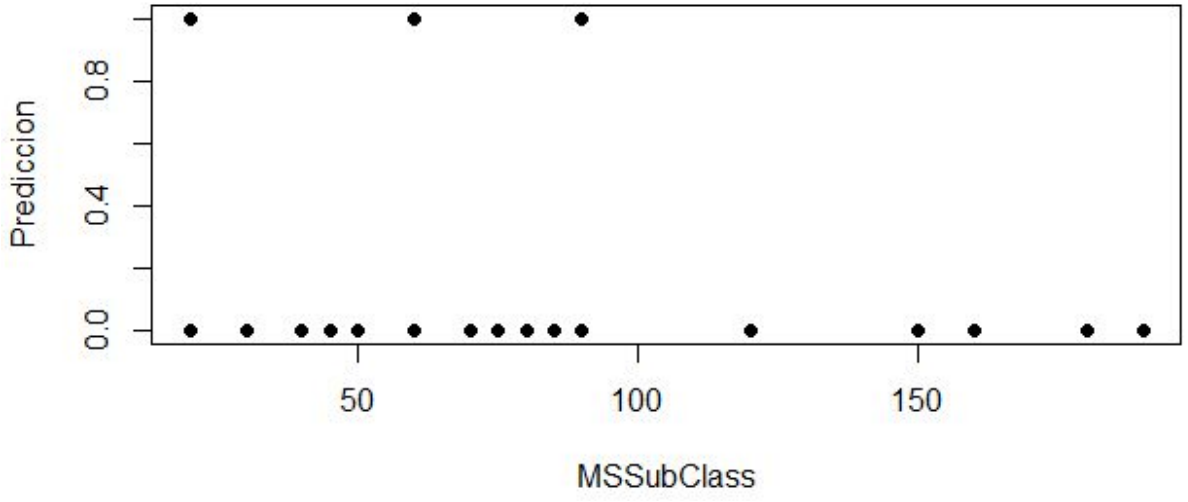
Para aplicar Regresión Logística se utilizaron las mismas variables que el modelo de Naive Bayes, las cuales eran: MSSubClasss, OverallCond, YearBuilt, BsmtFinSF1, X2ndFlrSF, BsmtFullBath, BedroomAbvGr y SceenPorch.
Para este modelo solo era necesario saber si la casa era Cara o no, por lo que los límites de clasificación cambiaron; si el precio era mayor a $200,000 es Cara, si es menor es Económica. Después de determinar la clasificación, se agregaron dos columnas dummys para indicar con 0 y 1 a qué clase pertenece cada dato. Esto sirve para poder aplicar el modelo de regresión logística. En los pasos anteriores, se juntó todo el dataset de training y test. Cuando ya se agregaron las columnas dummy se volvió a separar el dataset en training (70%) y test (30%).

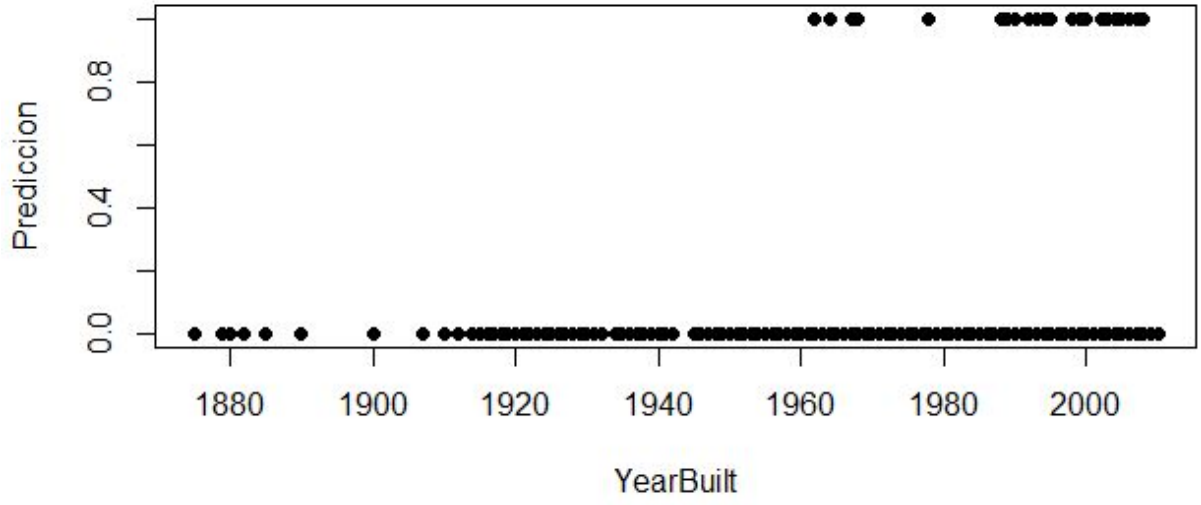
X2ndFlrSF	BsmtFullBath	BedroomAbvGr	ScreenPorch	Class	dataCara	dataEconomica
854	1	3	0	Cara	1	0
0	0	3	0	Economica	0	1
866	1	3	0	Cara	1	0
756	1	3	0	Economica	0	1
1053	1	4	0	Cara	1	0
566	1	1	0	Economica	0	1
0	1	3	0	Cara	1	0
983	1	3	0	Cara	1	0
752	0	2	0	Economica	0	1
0	1	2	0	Economica	0	1
0	1	3	0	Economica	0	1
1142	1	4	0	Cara	1	0

```
> modelo

Call:  glm(formula = dataCara ~ ., family = binomial(), data = train[,
    c(1:8, 10)], maxit = 100)

Coefficients:
 (Intercept)    MSSubClass   OverallCond     YearBuilt    BsmtFinSF1
  -5.999e+01    -8.471e-03     1.259e-01     2.779e-02     7.381e-04
  X2ndFlrSF  BsmtFullBath  BedroomAbvGr   ScreenPorch
   9.975e-04     7.958e-02     8.185e-01     2.640e-03

Degrees of Freedom: 2040 Total (i.e. Null);  2032 Residual
Null Deviance:      1956
Residual Deviance: 1573         AIC: 1591
```

Gráficamente, algunos ejemplos del modelo son:

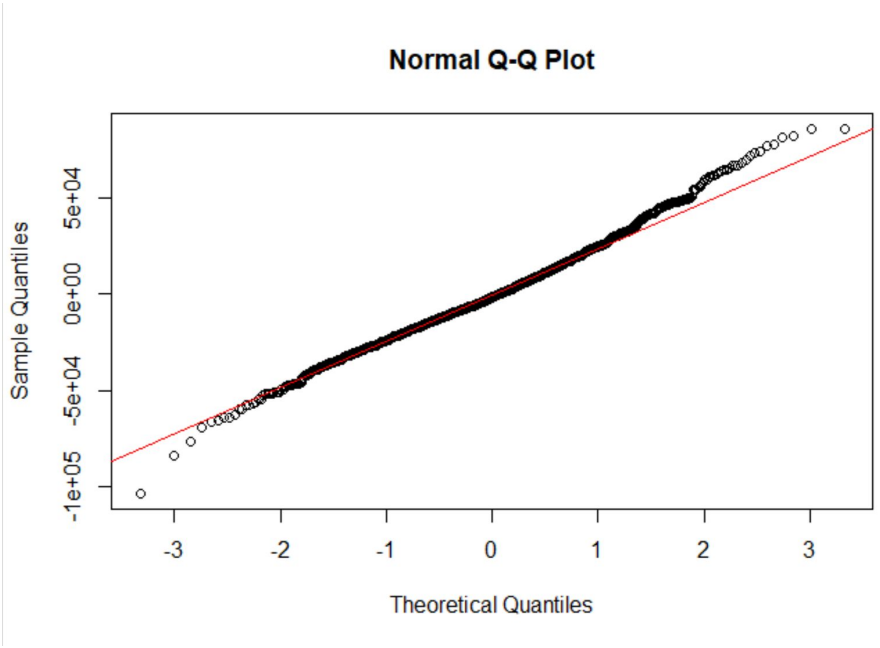


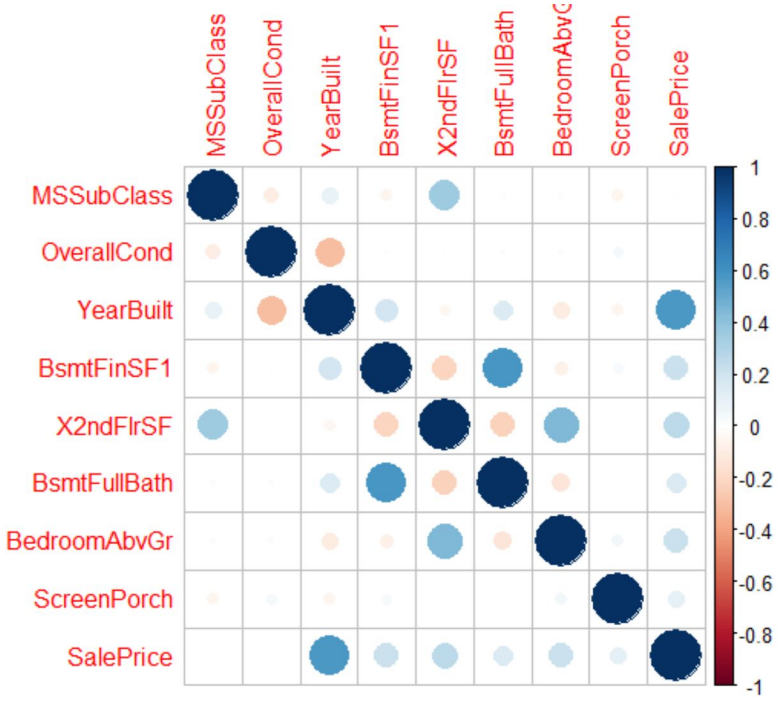




Análisis de las variables

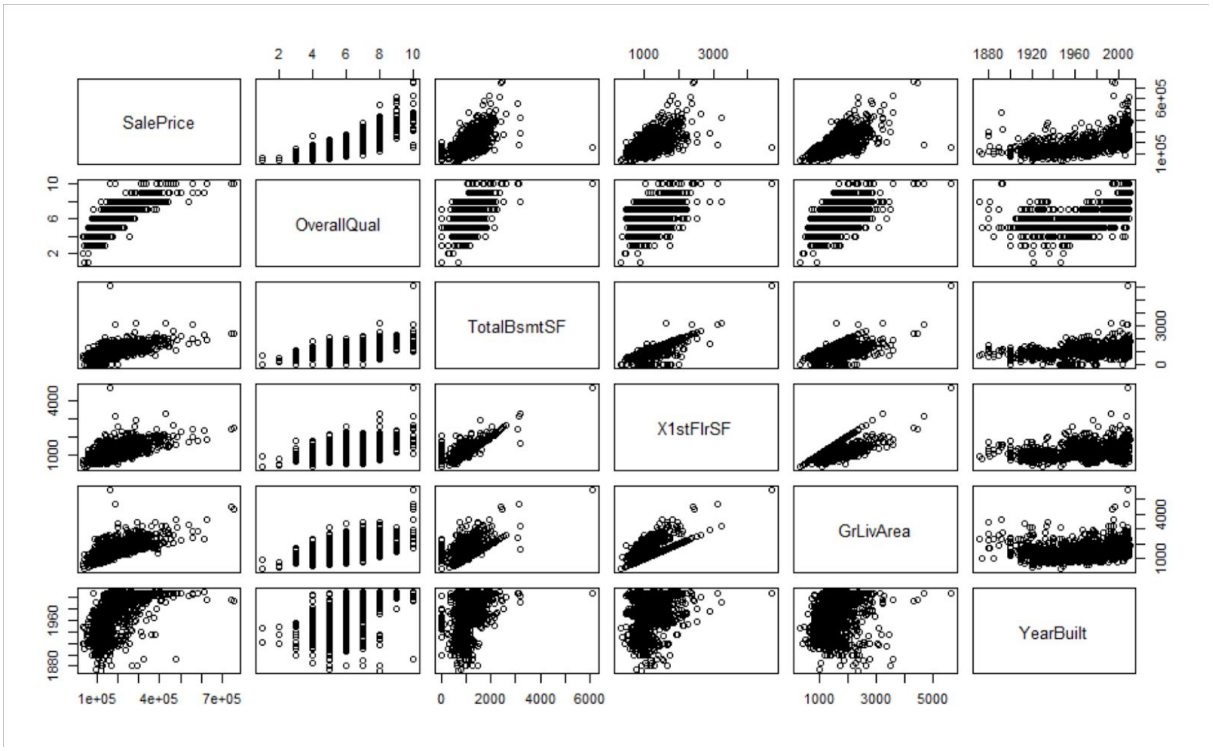
Las pruebas de normalidad sobre los datos produjeron las siguientes gráficas. Primero, se observa que las variables son normales.





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

Al aplicar el modelo de Regresión Lineal en el conjunto de prueba, se dieron los siguientes resultados, una lista conteniendo 0 y 1 según haya considerado su clasificación. El modelo clasificó la mayoría de casas como Económicas. Más adelante se analiza el accuracy del modelo en la predicción.

```
> prediccion
  1   5   6  10  12  13  19  24  27  28  31  34
  0   1   0   0   1   0   0   0   0   0   0   0
 35  37  40  41  43  47  48  49  50  51  52  54
  0   0   0   0   0   0   0   0   0   0   0   0
 55  62  63  66  68  83  85  87  88  89  95 102
  0   0   0   1   0   0   0   0   0   0   0   0
103 107 108 119 121 122 132 135 138 140 145 147
  0   0   0   1   0   0   0   0   0   0   0   0
149 150 151 154 156 159 161 163 168 170 177 183
  0   0   0   0   0   0   0   0   0   0   1   0
184 188 191 197 203 206 207 209 214 221 223 228
  0   0   0   0   0   0   0   0   0   0   0   0
232 234 235 240 241 244 245 250 253 255 257 263
```

Matriz de confusión

Para determinar la eficiencia de la regresión logística, se utilizó una matriz de confusión que compara cuánto se equivocó de los datos reales. Se obtuvo un accuracy del **83.6%**, indicando que el modelo predice muy bien la clasificación de las casas y no hace overfitting.
Se puede observar que lo que mejor predice el modelo fueron las casas Económicas como Económicas. Se confundió más en decir que ciertas casas son Caras cuando realmente eran Económicas.

```
> confusionMatrix(as.factor(test$dataCara),as.factor(prediccion))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 698   6
         1 138  34

               Accuracy : 0.8356
                 95% CI : (0.8094, 0.8596)
    No Information Rate : 0.9543
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2664

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

            Sensitivity : 0.8349
            Specificity : 0.8500
         Pos Pred Value : 0.9915
         Neg Pred Value : 0.1977
             Prevalence : 0.9543
         Detection Rate : 0.7968
   Detection Prevalence : 0.8037
      Balanced Accuracy : 0.8425

       'Positive' Class : 0
```

Comparación de modelos

Si se compara el modelo de Regresión Logística con el del árbol de clasificación, se puede concluir que son distintos en cuanto a resultados y desempeño; la diferencia de accuracy en la matriz de confusión fue de 14.13%. Si se compara el modelo de Regresión Logística con el del Naive Bayes, se puede concluir que son distintos también; la diferencia de accuracy en la matriz de confusión fue de 13.66%.
Además de tener un accuracy mejor -el modelo de regresión logística-, las líneas de código son menos y por ende se tarda menos en procesar. Sin embargo, la regresión logística solo nos permite clasificar de forma binaria (pertenece o no). Si necesitamos más de tres clasificadores, no es útil y se recomienda usar Naive Bayes.