Universidad del Valle de Guatemala
Minería de Datos
Sección 10

Hoja de Trabajo 5

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

Guatemala, 18 de marzo del 2024

1. Use los mismos conjuntos de entrenamiento y prueba que utilizó en las dos hojas anteriores.
2. Elabore un modelo de regresión usando bayes ingenuo (naive bayes), el conjunto de entrenamiento y la variable respuesta SalesPrice. Prediga con el modelo y explique los resultados a los que llega. Asegúrese que los conjuntos de entrenamiento y prueba sean los mismos de las hojas anteriores para que los modelos sean comparables.

```
> prediction
  [1] 142500 205000 132500 142000 112500 142500 142500 112500 215000 203000 159000 208900 189000 223500 116000 113000 131500
 [18] 192000 142500 142500 149900 124000 142500 223500 208900 112000 219500 114500 161500 168500 203000 224900 133900 113000
 [35] 174000 142500 159500 213500 219500 174000 203000 201000 203000 117500 208900 228500 142500 142500 113000 167000 152000
 [52] 112500 112000 113000 205000 262500 180500 156000 112500 131000 189000 205000 203000 142500 131500 142000 188000 203000
 [69] 213500 192000 137000 187500 167000 189000 153000 186500 211000 208900 142500 192000 142000 137000 203000 203000 116000
 [86] 142500 112000 142500 186500 116000 174000 188000 161500 219500 131000 125500 125500 211000 156000 131000 187500 112500
[103] 208900 186500 117500 112500 118000 116000 208900 205000 189000 186500 128500 203000 189000 266000 167000 142500 112500
[120] 189000 118000 142000 167000 174000 148500 159000 142500 167000 187500 237000 131500 167000 112000 131000 178000 221000
[137] 203000 142000 125500 156000 117000 203000 144500 112500 186500 208900 142500 223500 142500 203000 208900 228500 129500
[154] 219500 203000 186500 136500 177500 189000 120500 175900 112500 142500 224900 219500 196000 189000 116000 224900 167000
[171] 159500 142500 186500 186500 131000 142500 112000 132000 142500 113000 180500 169000 112000 142500 131000 168000 142500
[188] 219500 132500 197000 208900 112000 142500 142000 197000 232000 152000 112500 142500 205000 142500 197000 219500 203000
[205] 124000 188000 142000 136500 208900 159500 197000 136500 136500 132000 174000 203000 175900 113000 112000 196000 239000
[222] 167000 212000 142500 197000 196000 131000 142500 112500 142000 203000 186500 224900 132000 125500 132500 142500 168000
[239] 112500 189000 174000 113000 189000 208900 215000 113000 188000 167500 112500 113000 113000 187500 171000 112500 187500
[256] 142500 197000 113000 213500 142500 186500 186500 112000 208900 197000 203000 186500 142500 136500 112500 203000 117500
[273] 197000 208900 203000 223500 142500 134500 197000 203000 142500 113000 224900 180500 112500 123000 119500 256000 142500
[290] 224900 131000 208900 203000 174000 133900 186500 174000 159500 112500 191000 186500 224900 168000 136500 186500 117500
[307] 196000 125500 113000 189000 136500 116000 124000 134000 142500 142500 208900 148500 228500 112500 116000 203000 203000
```

3. Analice los resultados del modelo de regresión. ¿Qué tan bien le fue prediciendo?
MSE: 1.98
RMSE: 1.41
R^2: -0.99

A partir del coeficiente de determinación obtenido, se puede decir que el modelo de regresión generado es totalmente inadecuado para 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. Probar con otros modelos.
2. Validación cruzada.
3. Desnormalizar los datos, puesto que el uso de decimales perjudica ya que Naive Bayes es utilizado para modelos de clasificación.

MSE: 974292879
RMSE: 31213.66
R^2: 0.35

A partir del coeficiente de determinación obtenido, se puede decir que el árbol de regresión logra explicar solamente el 35% 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. Probar con otros modelos.
2. Validación cruzada.

4. Compare los resultados con el modelo de regresión lineal y el árbol de regresión que hizo en las hojas pasadas. ¿Cuál funcionó mejor?
El coeficiente de determinación del modelo de regresión lineal univariado de la hoja #3 es de 0.5, mientras que el coeficiente de determinación del modelo de regresión lineal multivariado es de 0.79. Finalmente, el coeficiente de determinación del modelo de regresión de la hoja anterior es de 0.62. Por lo tanto se puede afirmar que el modelo de regresión generado en esta hoja es peor que todos los modelos anteriores.

5. Haga un modelo de clasificación, use la variable categórica que hizo con el precio de las casas (barata, media y cara) como variable respuesta.
La variable de respuesta es la misma que fue generada en la hoja anterior.

6. Utilice los modelos con el conjunto de prueba y determine la eficiencia del algoritmo para predecir y clasificar.
7. Haga un análisis de la eficiencia del modelo de clasificación usando una matriz de confusión. 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         81        12        117
  Economica     3        71         35
  Intermedia    2         0          2

Overall Statistics

               Accuracy : 0.4768
                 95% CI : (0.4212, 0.5328)
    No Information Rate : 0.4768
    P-Value [Acc > NIR] : 0.5219

                  Kappa : 0.2874

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

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.9419           0.8554          0.012987
Specificity               0.4557           0.8417          0.988166
Pos Pred Value            0.3857           0.6514          0.500000
Neg Pred Value            0.9558           0.9439          0.523511
Prevalence                0.2663           0.2570          0.476780
Detection Rate            0.2508           0.2198          0.006192
Detection Prevalence      0.6502           0.3375          0.012384
Balanced Accuracy         0.6988           0.8485          0.500576
```

Según los resultados de la matriz de confusión, el modelo tuvo una exactitud de 0.48. Tuvo mayor cantidad de errores al predecir una casa como categoría intermedia. Esto concuerda con la observación realizada en la hoja anterior, en la que se veía una clara tendencia en clasificaciones de tipo “intermedia”. En general, el modelo parece tener un rendimiento bajo con una exactitud del 48% y un valor de kappa de 0.2874. La sensibilidad es más alta para la clase "Cara", lo que indica que el modelo es mejor identificando esta clase correctamente en comparación con las otras. La especificidad es alta para la clase "Intermedia", lo que sugiere que el modelo es bueno para identificar las no intermedias.

8. Analice el modelo. ¿Cree que pueda estar sobre ajustado?

```
Confusion Matrix and Statistics

           Reference
Prediction   Cara Economica Intermedia
  Cara         82         1         14
  Economica   103       184        345
  Intermedia    7         2         13

Overall Statistics

               Accuracy : 0.3715
                 95% CI : (0.3368, 0.4072)
    No Information Rate : 0.4953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.154

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

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.4271           0.9840           0.03495
Specificity               0.9732           0.2057           0.97625
Pos Pred Value            0.8454           0.2911           0.59091
Neg Pred Value            0.8318           0.9748           0.50754
Prevalence                0.2557           0.2490           0.49534
Detection Rate            0.1092           0.2450           0.01731
Detection Prevalence      0.1292           0.8415           0.02929
Balanced Accuracy         0.7001           0.5948           0.50560
```

Al comparar la exactitud del modelo utilizando los datos de entrenamiento con la exactitud utilizando los datos de prueba, es posible inducir que que el modelo no está sobre ajustado y en su lugar, podría estar sub ajustado.

9. Haga un modelo usando validación cruzada, compare los resultados de este con los del modelo anterior. ¿Cuál funcionó mejor?

```
Confusion Matrix and Statistics

           Reference
Prediction   Cara Economica Intermedia
  Cara         62         0         14
  Economica     2        68         49
  Intermedia   23         7         98

Overall Statistics

               Accuracy : 0.7059
                 95% CI : (0.6529, 0.755)
    No Information Rate : 0.4985
    P-Value [Acc > NIR] : 3.037e-14

                  Kappa : 0.55

 Mcnemar's Test P-Value : 8.712e-08

Statistics by Class:

                     Class: Cara Class: Economica
Sensitivity               0.7126           0.9067
Specificity               0.9407           0.7944
Pos Pred Value            0.8158           0.5714
Neg Pred Value            0.8988           0.9657
Prevalence                0.2693           0.2322
Detection Rate            0.1920           0.2105
Detection Prevalence      0.2353           0.3684
Balanced Accuracy         0.8267           0.8505
                     Class: Intermedia
Sensitivity                     0.6087
Specificity                     0.8148
Pos Pred Value                  0.7656
Neg Pred Value                  0.6769
Prevalence                      0.4985
Detection Rate                  0.3034
Detection Prevalence            0.3963
Balanced Accuracy               0.7118
```

Según la matriz de confusión, este modelo posee una exactitud de 0.71 y un valor de kappa de 0.55. Funcionó mejor que el modelo de clasificación realizado anteriormente.

10. Tanto para los modelos de regresión como de clasificación,pruebe con varios valores de los hiperparámetros, use el mejor modelo del tunneo, ¿Mejoraron los modelos? Explique

Modelo de clasificación

```
Confusion Matrix and Statistics

            Reference
Prediction   Cara Economica Intermedia
  Cara         58         1         46
  Economica     4        81         42
  Intermedia   11         4         76

Overall Statistics

               Accuracy : 0.6656
                 95% CI : (0.6113, 0.7169)
    No Information Rate : 0.5077
    P-Value [Acc > NIR] : 6.903e-09

                  Kappa : 0.5074

 Mcnemar's Test P-Value : 8.025e-12

Statistics by Class:

                     Class: Cara Class: Economica Class: Intermedia
Sensitivity               0.7945           0.9419            0.4634
Specificity               0.8120           0.8059            0.9057
Pos Pred Value            0.5524           0.6378            0.8352
Neg Pred Value            0.9312           0.9745            0.6207
Prevalence                0.2260           0.2663            0.5077
Detection Rate            0.1796           0.2508            0.2353
Detection Prevalence      0.3251           0.3932            0.2817
Balanced Accuracy         0.8033           0.8739            0.6845
```

La exactitud obtenida tras modificar el parámetro de laplace en el modelo de clasificación es de 0.67, por lo que el modelo mejoró respecto al modelo original, en el que la exactitud era de 0.48.

Modelo de regresión

```
> tune_grid <- expand.grid(
+   usekernel = c(TRUE, FALSE),
+   fL = seq(0, 1, length = 10), # Explora valores entre 0 y 1
+   adjust = 1 # Puedes variar esto si 'usekernel' es TRUE
+ )
>
> ct<-trainControl(method = "cv",number=10, verboseIter=T)
> modeloCaret<-train(SalePrice~.,data=train_copy,method="nb",trControl = ct,tuneGrid=tune_grid)
Error: wrong model type for regression
```

No se pudo realizar el tuneo del parámetro laplace, puesto que la librería caret identifica que la variable de respuesta es numérica. El utilizar naiveBayes directamente no produce este error, puesto que los datos numéricos son considerados como categóricos, por lo que para la evaluación de dicho modelo únicamente se casteó la predicción a un valor entero.

11. Compare la eficiencia del algoritmo con el resultado obtenido con el árbol de decisión (el de clasificación) y el modelo de random forest que hizo en la hoja pasada. ¿Cuál es mejor para predecir? ¿Cuál se demoró más en procesar?

Comparando todos los modelos realizados hasta ahora en esta hoja y la hoja anterior, el mejor modelo para predecir es el generado con Random Forest con una exactitud de 0.78. Sin embargo, la exactitud del modelo generado a partir de validación cruzada con Naive Bayes, posee un valor cercano al de Random Forest, por lo que también se puede considerar como un buen modelo para predecir si una casa es considerada cara, económica o intermedia. El modelo generado a partir de validación cruzada en esta hoja demoró más que el modelo generado con Random Forest.