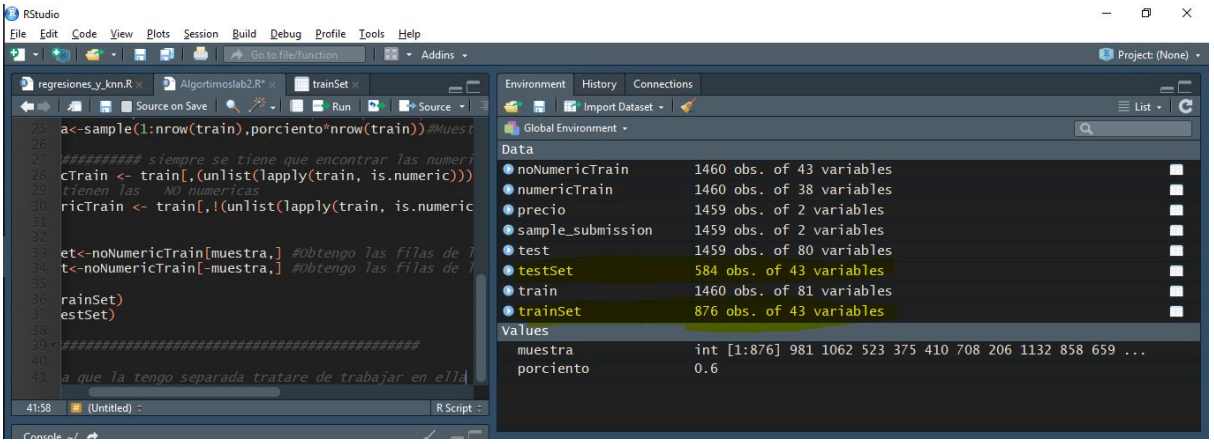
Universidad del Valle de Guatemala
Data Science
Lynette Pérez

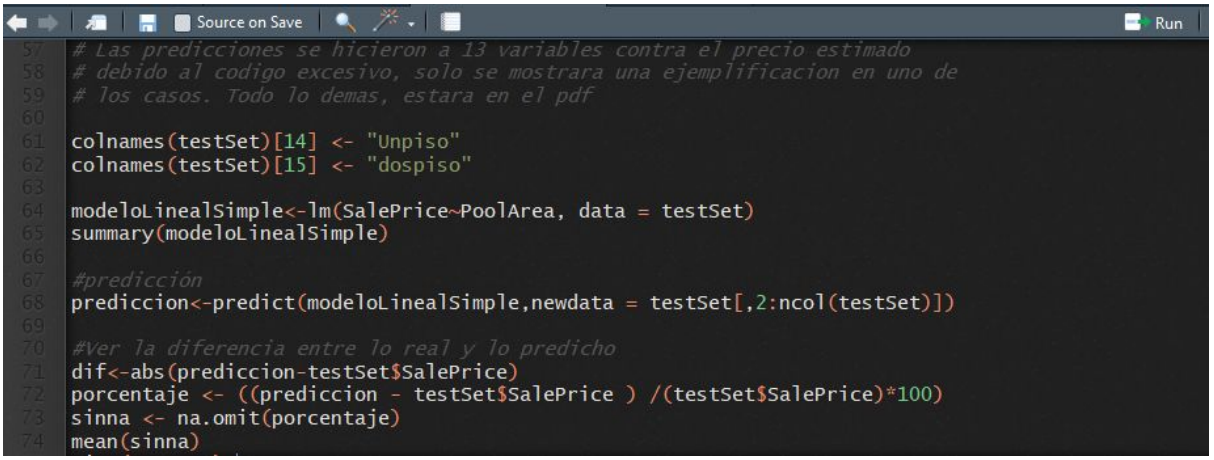
Lab2 : Aprendizaje de algoritmos

Jorge Eduardo Súchite
Carnet 15293

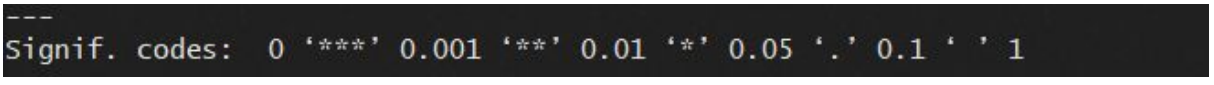
1. **Divida el set de datos de entrenamiento que le provee kaggle en 2 conjuntos, entrenamiento (60%) y prueba (40%). Las filas que van a cada subconjunto se seleccionan aleatoriamente.**

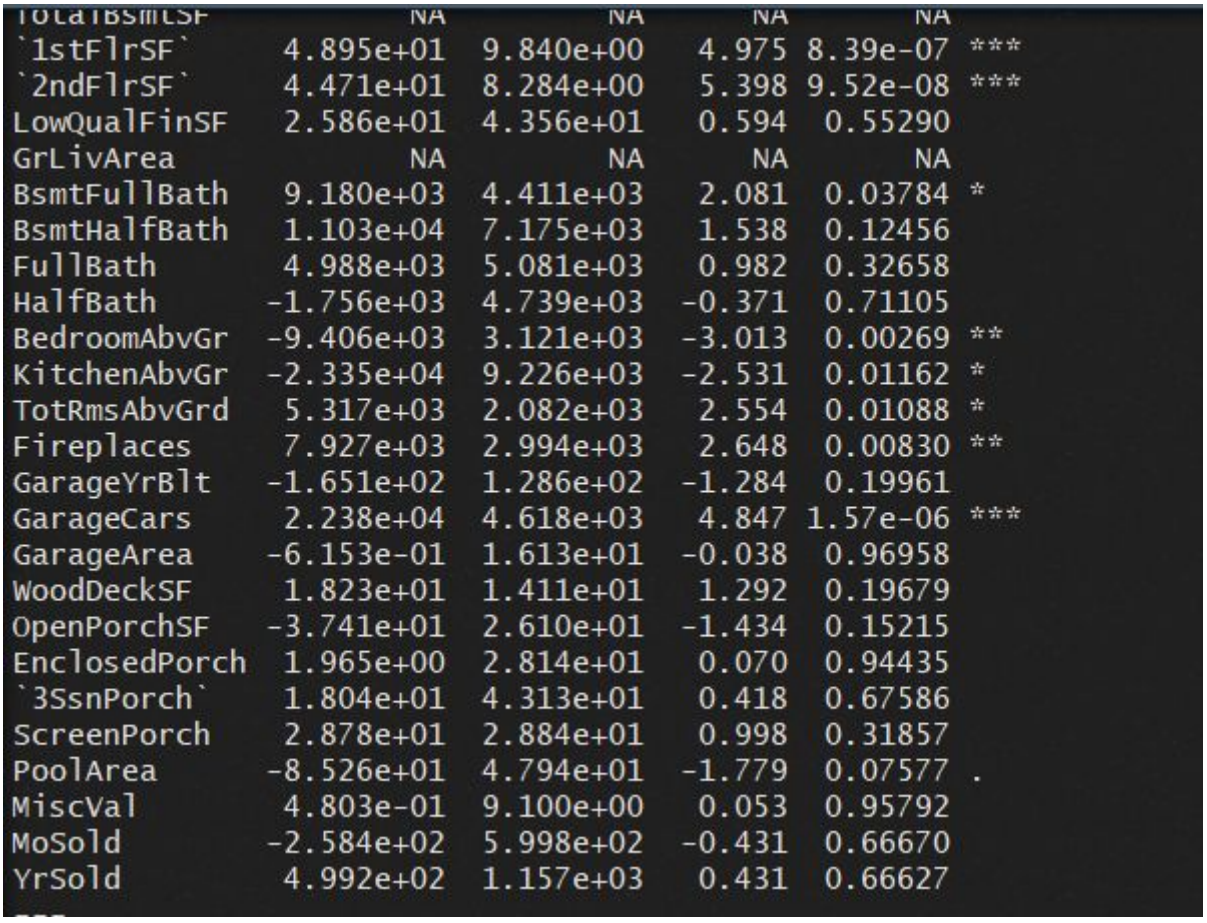


2. **Haga un modelo de regresión lineal para predecir el precio de las casas. Como ya hizo un análisis exploratorio del conjunto de datos, explique la selección de variables con los que hizo el modelo.**



3. **Haga un análisis del modelo generado, ¿Cuáles son las variables significativas? ¿Explica o no la variabilidad de los datos? Si considera necesario redefinir las variables del modelo, hágalo y explique las causas.**





TotalBsmtSF	NA	NA	NA	NA	
`1stFlrSF`	4.895e+01	9.840e+00	4.975	8.39e-07	***
`2ndFlrSF`	4.471e+01	8.284e+00	5.398	9.52e-08	***
LowQualFinSF	2.586e+01	4.356e+01	0.594	0.55290	
GrLivArea	NA	NA	NA	NA	
BsmtFullBath	9.180e+03	4.411e+03	2.081	0.03784	*
BsmtHalfBath	1.103e+04	7.175e+03	1.538	0.12456	
FullBath	4.988e+03	5.081e+03	0.982	0.32658	
HalfBath	-1.756e+03	4.739e+03	-0.371	0.71105	
BedroomAbvGr	-9.406e+03	3.121e+03	-3.013	0.00269	**
KitchenAbvGr	-2.335e+04	9.226e+03	-2.531	0.01162	*
TotRmsAbvGrd	5.317e+03	2.082e+03	2.554	0.01088	*
Fireplaces	7.927e+03	2.994e+03	2.648	0.00830	**
GarageYrBlt	-1.651e+02	1.286e+02	-1.284	0.19961	
GarageCars	2.238e+04	4.618e+03	4.847	1.57e-06	***
GarageArea	-6.153e-01	1.613e+01	-0.038	0.96958	
WoodDeckSF	1.823e+01	1.411e+01	1.292	0.19679	
OpenPorchSF	-3.741e+01	2.610e+01	-1.434	0.15215	
EnclosedPorch	1.965e+00	2.814e+01	0.070	0.94435	
`3SsnPorch`	1.804e+01	4.313e+01	0.418	0.67586	
ScreenPorch	2.878e+01	2.884e+01	0.998	0.31857	
PoolArea	-8.526e+01	4.794e+01	-1.779	0.07577	.
MiscVal	4.803e-01	9.100e+00	0.053	0.95792	
MoSold	-2.584e+02	5.998e+02	-0.431	0.66670	
YrSold	4.992e+02	1.157e+03	0.431	0.66627	

Significancia	**Variable**
0	MSSubClass OverallQual 1stfFlrSF 2ndFlrSF GarageCars
0.001	YearBuilt BedroomAbvGr Fireplaces
0.01	LotFrontage LotArea MasVnrArea BsmtFullBath KitchenAbvGr TotRmsAbvGrd
0.05	OverallCond PoolArea

0.1	YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF LowQualFinSF BsmtHalfBath FullBath HalfBath GarageArea WoodDeckSF OpenPorchSF EnclosedPorch `3SsnPorch` ScreenPorch
1	**Ninguna**

Correlación indica cuáles de las variables estuvieron más cercanas al 1 en la correlación lineal. Por ende, estas serán las que influyen al momento de otorgar un precio a cada casa. Elegí todas las que sean mayores o iguales a 0.1 y las más importantes que se considera de todas las secciones de significancia.

4. **Compare el precio que predijo el algoritmo con el que ya se conoce, explique la efectividad del algoritmo definiendo una diferencia mínima. Explique la elección del número que marca la diferencia.**

Este algoritmo es mejor debido a que en el laboratorio anterior. Mis variables para decidir el precio de una casa eran como la calidad de la cocina ; el aire acondicionado, cosas nada que ver y en este es más puntual y lógico

5. **Haga un modelo de KNN (K nearest neighbors). Explique la elección del parámetro k.**

```
          Reference
Prediction   0   2   3   4   5
         0 311  42   0   0   0
         2  91 109   0   0   0
         3   0  25   1   0   0
         4   0   3   1   0   0
         5   0   0   1   0   0

Overall Statistics

               Accuracy : 0.7209
                 95% CI : (0.6826, 0.7569)
    No Information Rate : 0.6884
    P-Value [Acc > NIR] : 0.04813

                  Kappa : 0.417
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.7736   0.6089 0.333333       NA       NA
Specificity            0.7692   0.7753 0.956971 0.993151 0.998288
Pos Pred Value         0.8810   0.5450 0.038462       NA       NA
Neg Pred Value         0.6061   0.8177 0.996416       NA       NA
Prevalence             0.6884   0.3065 0.005137 0.000000 0.000000
Detection Rate         0.5325   0.1866 0.001712 0.000000 0.000000
Detection Prevalence   0.6045   0.3425 0.044521 0.006849 0.001712
Balanced Accuracy      0.7714   0.6921 0.645152       NA       NA
```

72% de efectividad

Se utilizó un k de 23 puesto que la raíz de toda la data resultaba en 24.16 y por números más convenientes se eligió el número 23 porque los datos eran impares.

68% de los datos no los tomó en cuenta

7. Vuelva a ejecutar los modelos usando validación cruzada. Compare los resultados obtenidos.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   2   3   4   5
         0 311  42   0   0   0
         2  91 109   0   0   0
         3   0  25   1   0   0
         4   0   3   1   0   0
         5   0   0   1   0   0

Overall Statistics

               Accuracy : 0.7209
                 95% CI : (0.6826, 0.7569)
    No Information Rate : 0.6884
    P-Value [Acc > NIR] : 0.04813

                  Kappa : 0.417
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.7736   0.6089 0.333333       NA       NA
Specificity            0.7692   0.7753 0.956971 0.993151 0.998288
Pos Pred Value         0.8810   0.5450 0.038462       NA       NA
Neg Pred Value         0.6061   0.8177 0.996416       NA       NA
Prevalence             0.6884   0.3065 0.005137 0.000000 0.000000
Detection Rate         0.5325   0.1866 0.001712 0.000000 0.000000
Detection Prevalence   0.6045   0.3425 0.044521 0.006849 0.001712
Balanced Accuracy      0.7714   0.6921 0.645152       NA       NA
>
```

Mismos resultados

8. Compare el rendimiento de ambos algoritmos y determine cuál de los dos logró predecir mejor el precio de las casas.

Comparo los dos y los dos me dieron lo mismo entonces no puedo afirmar que uno sea mejor que el otro.