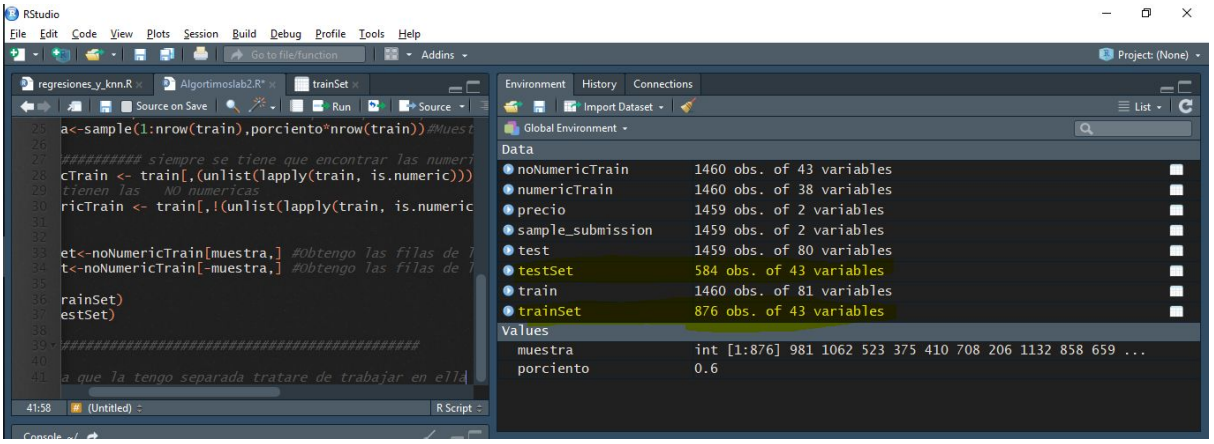
Universidad del Valle de Guatemala
Data Science
Lynette Pérez

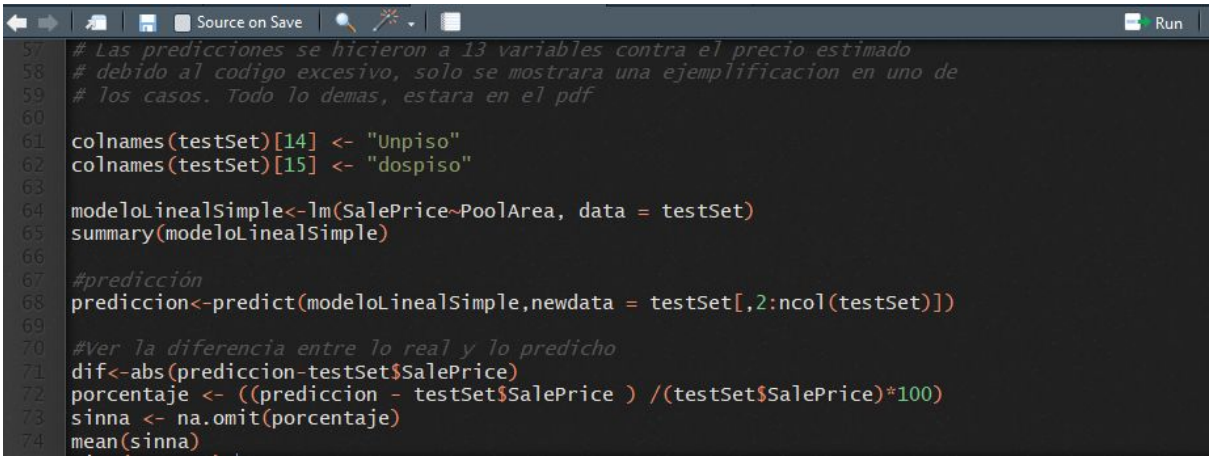
Lab2 : Aprendizaje de algoritmos

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.**

Variable	Procentaje de error
LotFrontage - Descartada	14.94732

LotArea - Descartada	14.5506
OverallQual - No descartada	4.248856
YearBuilt - Descartada	10.91941
YearRemodAdd- Descartada	10.84239
MasVnrArea- Descartada	13.69498
1stFlrSF - Descartada	10.60921
2ndFlrSF- Descartada	15.30983
GrLivArea - No descartada	9.067478
TotRmsAbvGrd - Descartada	12.25213
GarageYrBlt- Descartada	10.36411
GarageArea- No descartada	8.951427
GarageCars- No descartada	8.522686
WoodDeckSF- Descartada	15.15812
PoolArea- Descartada	16.3563
MSSubClass- Descartada	16.16761
BsmtFinSF1- Descartada	14.12561
BsmtFinSF2- Descartada	16.34418
BsmtUnfSF- Descartada	15.45138
TotalBsmtSF- Descartada	10.26352
FullBath- Descartada	10.50292
KitchenAbvGr- Descartada	16.06279
TotRmsAbvGrd- Descartada	12.25213
Fireplaces- Descartada	12.40693

El porcentaje de error de cada variable obtenida en la correlación lineal indica cuáles de las variables estuvieron más cercanas al 1 en la correlación lineal. Por ende, la calidad del material con la que está hecha la casa, el área del garage, el garage de carro y los pies cuadrados que tiene la casa tienen un relación lineal con el precio de las casas hablando de las variables cuantitativas de dicha casa.

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.**

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

Variable	**Procentaje de error**
LotFrontage - Descartada	
LotArea - Descartada	
OverallQual - No descartada	
YearBuilt - Descartada	
YearRemodAdd- Descartada	
MasVnrArea- Descartada	
1stFlrSF - Descartada	
2ndFlrSF- Descartada	
GrLivArea - No descartada	
TotRmsAbvGrd - Descartada	
GarageYrBlt- Descartada	
GarageArea- No descartada	
GarageCars- No descartada	
WoodDeckSF- Descartada	

PoolArea- Descartada	
MSSubClass- Descartada	
BsmtFinSF1- Descartada	
BsmtFinSF2- Descartada	
BsmtUnfSF- Descartada	
TotalBsmtSF- Descartada	
FullBath- Descartada	
KitchenAbvGr- Descartada	
TotRmsAbvGrd- Descartada	
Fireplaces- Descartada	

6. **Compare el precio que predijo el algoritmo (knn) con el que ya se conoce, explique la efectividad del algoritmo usando la diferencia mínima que definió en el ejercicio 3.**

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

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