Oscar Paredez
Guido Padilla
Diego Alvarez

Hoja de Trabajo 6.

Modelos de Regresión Logística

ACTIVIDADES

1. Cree una variable dicotómica por cada una de las categorías de la variable respuesta categórica que creó en hojas anteriores. Debería tener 3 variables dicotómicas (valores 0 y 1) una que diga si la vivienda es cara o no, media o no, económica o no.

```
    print('Mulitcoleanalidad de variable CARA')
    calculate_vif(df=df, features=['Cara', 'SalePrice', 'GrLivArea', 'OverallQual'])
```

```
Mulitcoleanalidad de variable CARA
```

```
    print('Mulitcoleanalidad de variable INTERMEDIA')
    calculate_vif(df=df, features=['Media', 'SalePrice', 'GrLivArea', 'OverallQual'])
```

```
Mulitcoleanalidad de variable INTERMEDIA
```

```
    print('Mulitcoleanalidad de variable ECONOMICA')
    calculate_vif(df=df, features=['Economica', 'SalePrice', 'GrLivArea', 'OverallQual'])
```

```
Mulitcoleanalidad de variable ECONOMICA
```

Oscar Paredez
Guido Padilla
Diego Alvarez

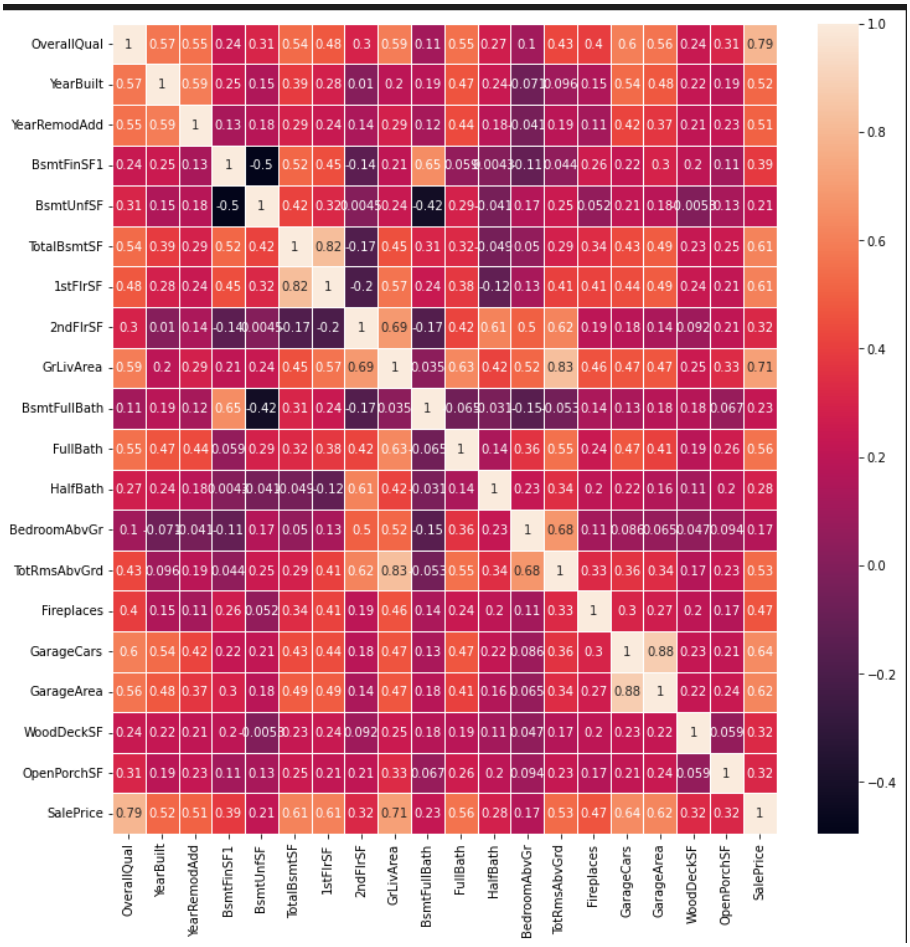
	VIF	Tolerance
Cara	1.300378	0.769007
SalePrice	4.264265	0.234507
GrLivArea	2.025104	0.493802
OverallQual	2.888613	0.346187

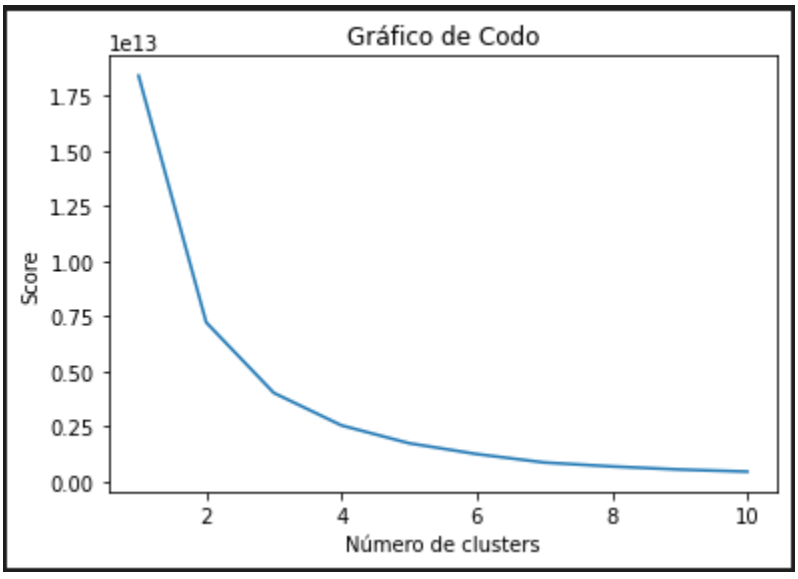
	VIF	Tolerance
Media	1.791565	0.558171
SalePrice	4.420400	0.226224
GrLivArea	2.029901	0.492635
OverallQual	2.686549	0.372225

	VIF	Tolerance
Economica	2.290518	0.436582
SalePrice	5.286043	0.189177
GrLivArea	2.037470	0.490805
OverallQual	2.707813	0.369302

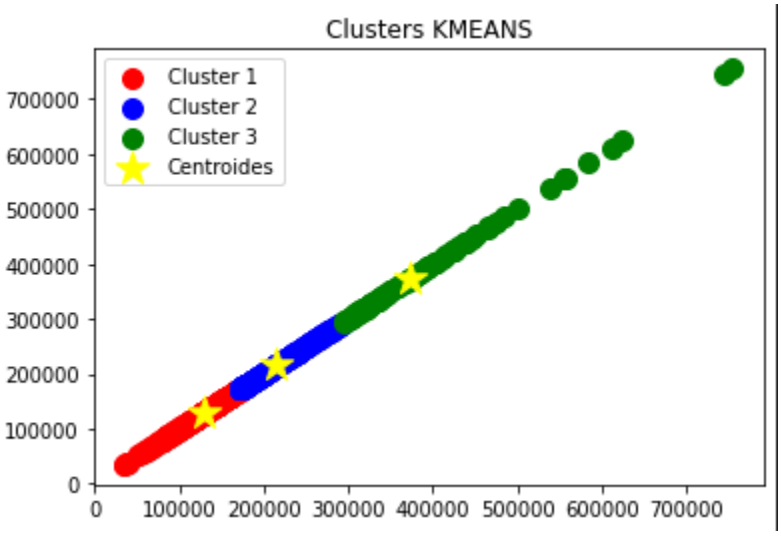
2. Use los mismos conjuntos de entrenamiento y prueba que utilizó en las hojas anteriores.

Oscar Paredez
Guido Padilla
Diego Alvarez





Oscar Paredez
Guido Padilla
Diego Alvarez




```
corrMatrix
```
✓ 0.1s

	OverallQual	YearBuilt	YearRemodAdd	BsmtFinSF1	BsmtUnfSF
OverallQual	1.000000	0.572323	0.550684	0.239666	0.308159
YearBuilt	0.572323	1.000000	0.592855	0.249503	0.149040
YearRemodAdd	0.550684	0.592855	1.000000	0.128451	0.181133
BsmtFinSF1	0.239666	0.249503	0.128451	1.000000	-0.495251
BsmtUnfSF	0.308159	0.149040	0.181133	-0.495251	1.000000
TotalBsmtSF	0.537808	0.391452	0.291066	0.522396	0.415360
1stFlrSF	0.476224	0.281986	0.240379	0.445863	0.317987
2ndFlrSF	0.295493	0.010308	0.140024	-0.137079	0.004469
GrLivArea	0.593007	0.199010	0.287389	0.208171	0.240257
BsmtFullBath	0.111098	0.187599	0.119470	0.649212	-0.422900
FullBath	0.550600	0.468271	0.439046	0.058543	0.288886
HalfBath	0.273458	0.242656	0.183331	0.004262	-0.041118
BedroomAbvGr	0.101676	-0.070651	-0.040581	-0.107355	0.166643

3. Elabore un modelo de regresión logística para conocer si una vivienda es cara o no, utilizando el conjunto de entrenamiento y explique los resultados a los que llega. Muestre el modelo gráficamente. El experimento debe ser reproducible por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.

Oscar Paredez
Guido Padilla
Diego Alvarez

```
Matriz de confusión para detectar casas caras
 [[435   2]
 [  1   0]]
Accuracy score:  0.9931506849315068
Presicion score:  0.9931506849315068
Recall score:  0.9931506849315068
F1 score:  0.9931506849315068
```

En este modelo de regresión obtuvimos un overfitting, el cual se dio por tener pocos datos, menos datos representativos o bien los datos son muy ruidosos.

4. Analice el modelo. Determine si hay multicolinealidad en las variables, y cuáles son las que aportan al modelo, por su valor de significación. Haga un análisis de correlación de las variables del modelo y especifique si el modelo se adapta bien a los datos. Explique si hay sobreajuste (overfitting) o no.

Mulitcoleanalidad de variable CARA

	VIF	Tolerance
Cara	1.300378	0.769007
SalePrice	4.264265	0.234507
GrLivArea	2.025104	0.493802
OverallQual	2.888613	0.346187

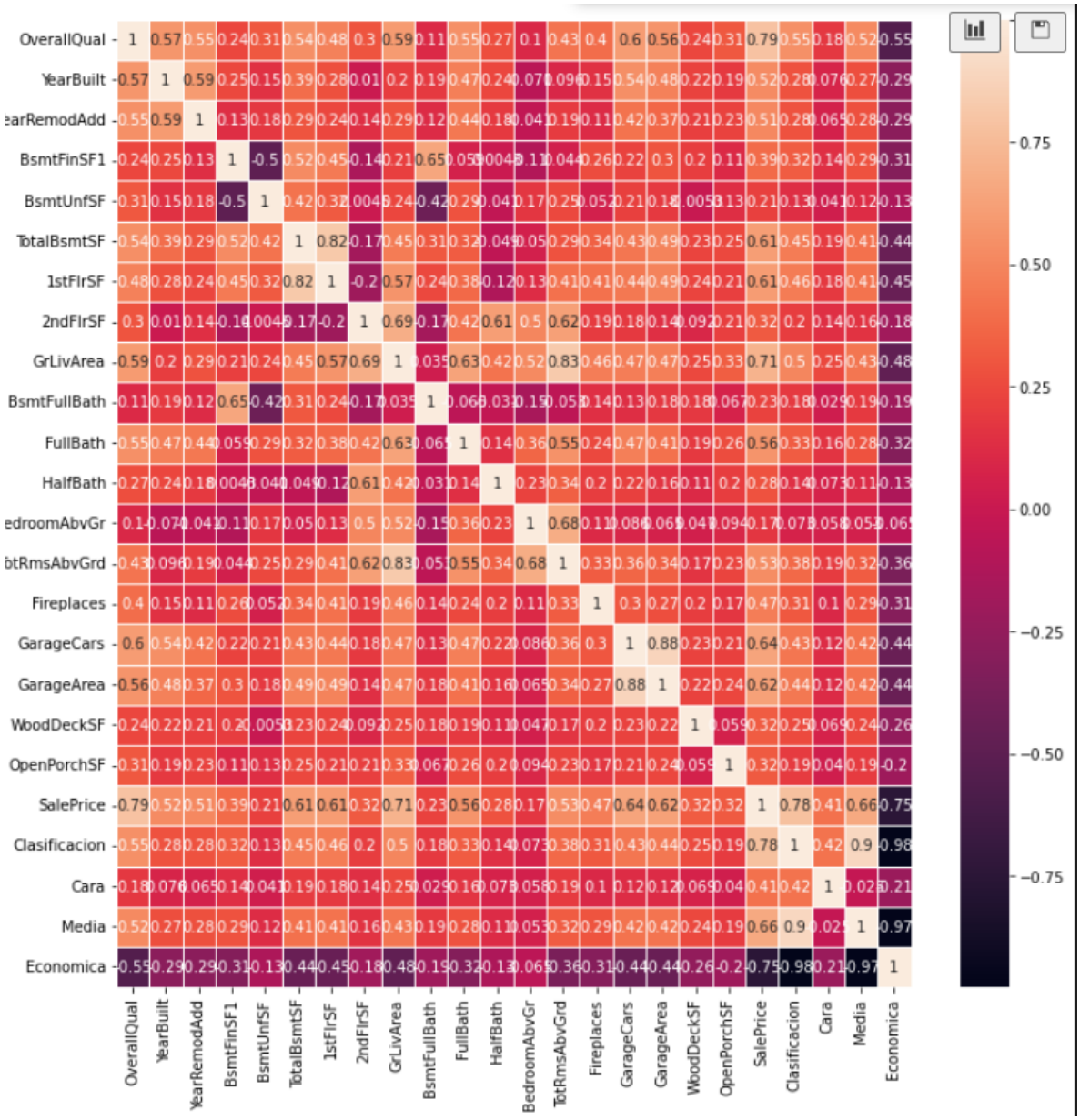
Mulitcoleanalidad de variable INTERMEDIA

	VIF	Tolerance
Media	1.791565	0.558171
SalePrice	4.420400	0.226224
GrLivArea	2.029901	0.492635
OverallQual	2.686549	0.372225

Oscar Paredez
Guido Padilla
Diego Alvarez

Mulitcoleanalidad de variable ECONOMICA

	VIF	Tolerance
Economica	2.290518	0.436582
SalePrice	5.286043	0.189177
GrLivArea	2.037470	0.490805
OverallQual	2.707813	0.369302



Oscar Paredez
Guido Padilla
Diego Alvarez

Se puede visualizar en el análisis de correlación la alta correlación y a la vez un alto índice de multicolinealidad con las variables de OverallQual, GrLivArea y SalePrice. Como se puede ver en el VIF previamente, SalePrice es la variable con más multicolinealidad, seguida de OverallQual y luego GrLivArea. Esto se replica en las tres variables analizadas (Economica, Media y Cara).

5. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar o predecir, en dependencia de las características de la variable respuesta.
6. Haga un análisis de la eficiencia del algoritmo 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.

Como se puede ver a continuación, se puede detectar que la efectividad más alta la tiene la matriz generada para detectar las casas caras, y la que menos efectividad tiene es la matriz para las casas intermedias.

```
Matriz de confusión para detectar casas Economicas
 [[ 38  10]
 [  3 387]]
Accuracy score:  0.9703196347031964
Presicion score:  0.9703196347031964
Recall score:  0.9703196347031964
F1 score:  0.9703196347031964
```

```
Matriz de confusión para detectar casas Intermedias
 [[384   7]
 [ 15  32]]
Accuracy score:  0.9497716894977168
Presicion score:  0.9497716894977168
Recall score:  0.9497716894977168
F1 score:  0.9497716894977168
```

```
Matriz de confusión para detectar casas caras
 [[435   2]
 [  1   0]]
Accuracy score:  0.9931506849315068
Presicion score:  0.9931506849315068
Recall score:  0.9931506849315068
F1 score:  0.9931506849315068
```

Oscar Paredez
Guido Padilla
Diego Alvarez

7. Cree otros dos modelos que determinen si una casa es barata o no, o intermedia o no. Repita para cada modelo, los pasos del 1- 6.

```
Matriz de confusión para detectar casas Intermedias
 [[384   7]
 [ 15  32]]
Accuracy score:  0.9497716894977168
Presicion score:  0.9497716894977168
Recall score:  0.9497716894977168
F1 score:  0.9497716894977168
```

```
Matriz de confusión para detectar casas caras
 [[435   2]
 [  1   0]]
Accuracy score:  0.9931506849315068
Presicion score:  0.9931506849315068
Recall score:  0.9931506849315068
F1 score:  0.9931506849315068
```

8. Compare la eficiencia de los 3 modelos que creó (uno para barata, otro para media y otro para cara) ¿Cuál se demoró más en procesar?¿Cuál se equivocó más?¿Cuál se equivocó menos?¿por qué?

Los resultados que se pueden observar en el inciso 6 demuestran la eficiencia de los 3 modelos utilizados. Se puede observar que la matriz de confusión para detectar las casas caras fue la que mejores puntajes tuvo en los distintos tipos de puntaje, con un 99.3%, seguido de los resultados para las casas económicas y finalmente las intermedias. La mas tardada por supuesto fue al que más datos entran en esa categoría, como lo es la clase económica. La que se equivocó más fue la de las casas intermedias y la que mejor resultado obtuvo fue la de las casas caras, posiblemente porque la cantidad de datos que entran en esta categoría es menor y las que pertenecen acá se encuentran más marcadas.

EVALUACIÓN

- **(25 puntos)** Análisis de los modelos generados. Recuerde explicar los razonamientos.
- **(25 puntos)** Análisis de las variables a incluir en los modelos. Pruebas de normalidad, correlación, etc. (Recuerde que las variables predictoras deberían ser las mismas para poder comparar)
- **(10 puntos)** Aplicación de los modelos al conjunto de prueba.

Oscar Paredez
Guido Padilla
Diego Alvarez

- **(20 puntos)** Matriz de confusión de cada modelo. Explicación de los resultados obtenidos
- **(20 puntos)** Comparación entre sí de los modelos generados.