

Tarea #: 2

Tema: Aprendizaje No supervisado y Regresión

Fecha entrega: 11:59 pm Septiembre 18 de 2023

Estudiante: Santiago Cárdenas Franco

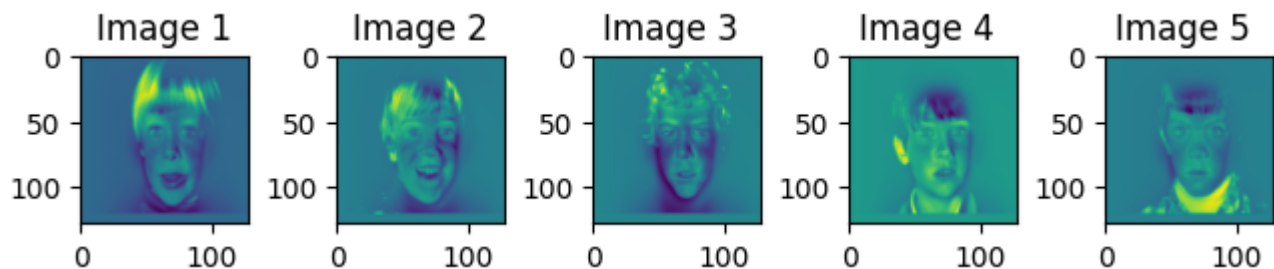
Objetivo: Aplicar los conceptos de PCA y regresión en datos reales.

Entrega: Crear una rama utilizando el mismo repositorio de la tarea 1, crear otra carpeta llamada tarea 2, solucionar el problema y crear un pull request sobre la master donde me debe poner como reviewer (entregas diferentes tienen una reducción de 0.5 puntos).

1 PCA (20%)

Cargar el data set de caras que está en la carpeta datos de la tarea 2 (ver notebook https://github.com/jdramirez/UCO_ML_AI/blob/master/src/notebook/PCA.ipynb):

1. Calcular la mean face. Que es la cara con el promedio de los pixeles y visualizarla.

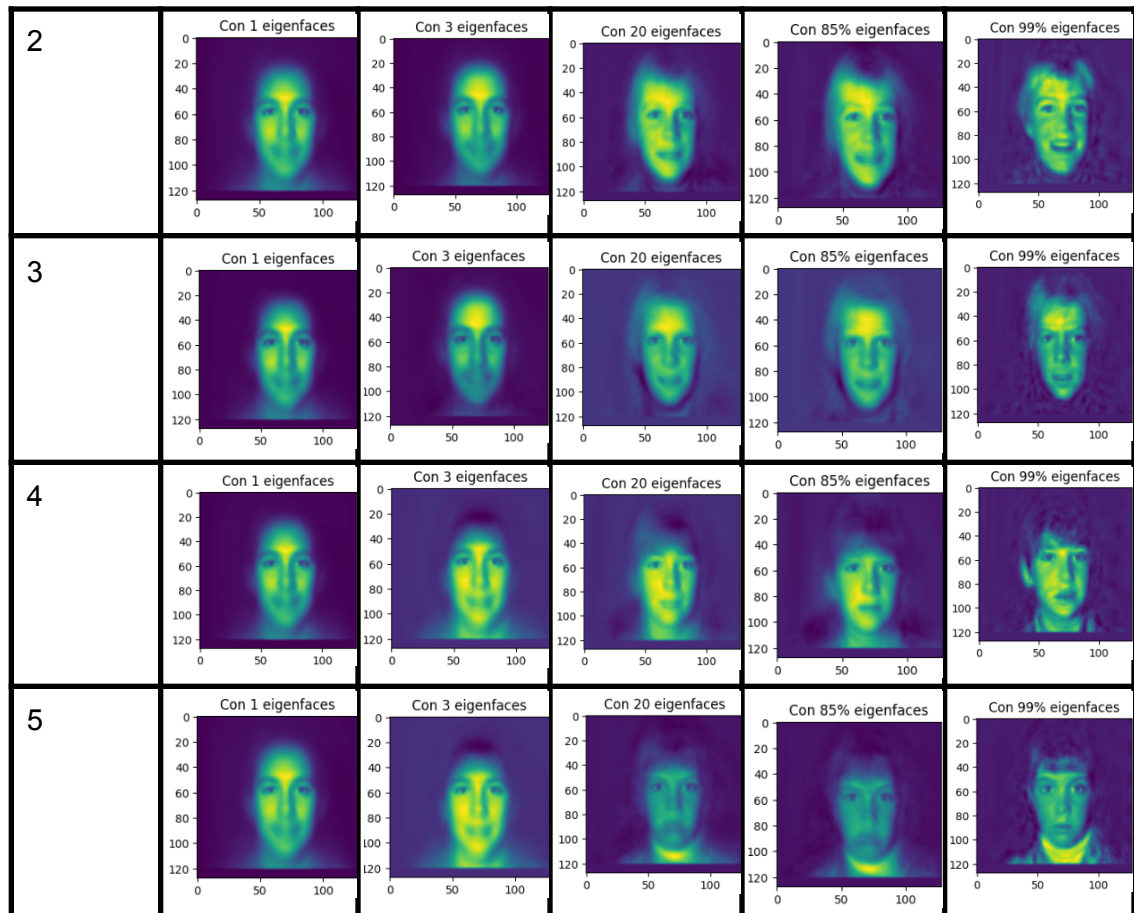


Se resaltan ciertas facciones de los rostros, por lo que se observa, se aprecia que calcular la mean face, puede observar rasgos importantes para cada cara.

2. Centrar los datos, utilizar PCA. ¿Cuántos componentes se deben utilizar para mantener el 90% de las características?. Crear una tabla para mostrar las primeras 5 caras utilizando, la mean face + los datos reconstruidos utilizando la primera componente, después con 5 componentes, después con las primeras 10 componentes, después con las componentes que explican el 90% de la varianza y por último con el número de componentes que tiene el 99% de la varianza.

¿Qué se puede concluir de los resultados?

Cara original	MeanFace + 1 comp	MeanFace + 3 comp	MeanFace + 20 comp	MeanFace + 85% comp	MeanFace + 99% comp
1	Con 1 eigenfaces 	Con 3 eigenfaces 	Con 20 eigenfaces 	Con 85% eigenfaces 	Con 99% eigenfaces



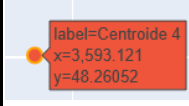
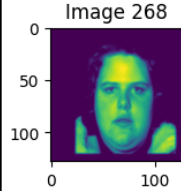
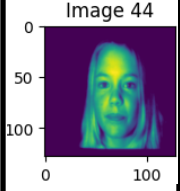
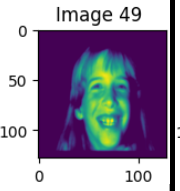
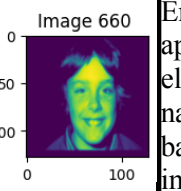
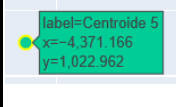
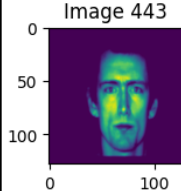
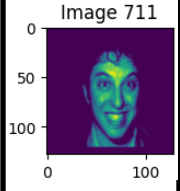
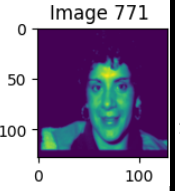
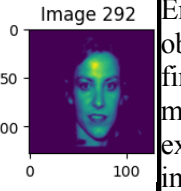
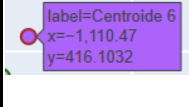
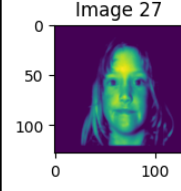
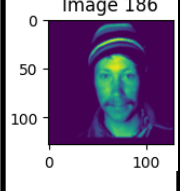
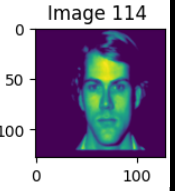
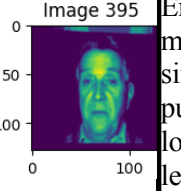
Se puede concluir que cuando se tiene una sola componente todas las imágenes tienden a una figura similar dando entender que cuando hay pocas componentes todas las imágenes comparten esa característica, dando a entender una mayor confusión y teniendo que buscar más datos para mejorar la figura, se aprecia que entre mayores componentes, mejores resultados. Se observa que en ciertas situaciones no es necesario tener el 100% de componentes para ver toda la información en este caso las imágenes.

2 K-means (20%)

Utilizar las 5 primeras componentes e implementar el algoritmo k-means sin librerías utilizando la distancia de valor absoluto (conocida como la norma 1), crear clase con métodos fit(aprender de los datos) y predict(predicir con los centroides el cluster de un nuevo dato).

1. Crear 7 clusters. Seleccione las 4 caras más cercanas al centroide de cada cluster, describa si son similares y porque están cerca una de la otra.

Cluster	Centroide	1ra Cara cercana al cluster	2da Cara cercana al cluster	3ra cara más cercana al cluster	4ta cara más cercana al cluster	¿Qué tienen en común?
1	<div> <div>label=Centroide 0</div> <div>x=1,313.406</div> <div>y=-1,041.487</div> </div>	<div>Image 463</div>	<div>Image 120</div>	<div>Image 126</div>	<div>Image 460</div>	Se aprecia que los rostros tienen varias similitudes, su forma de la cara es muy similar (Sobre todo para la imagen 463, 126 y 450), todos los rostros tienen forma cuadrática, en todas las imágenes se observan orejas grandes.
2	<div> <div>label=Centroide 1</div> <div>x=-1,395.038</div> <div>y=-715.6245</div> </div>	<div>Image 127</div>	<div>Image 250</div>	<div>Image 191</div>	<div>Image 399</div>	Entre las 4 En las imágenes se observa que los rostros presentan casi una forma ovalada, entre los tres se aprecian rasgos más pronunciados, como la nariz, en la imagen 250, 191 y 399 se puede observar una nariz amplia.
3	<div> <div>label=Centroide 2</div> <div>x=6,087.177</div> <div>y=4,120.84</div> </div>	<div>Image 648</div>	<div>Image 758</div>	<div>Image 759</div>	<div>Image 604</div>	Observando las 4 imágenes observamos que entre ellas tiene mayor brillo, se puede apreciar el fondo y sombras, además no tienen expresiones faciales muy exageradas, las imágenes 758, 759 y 604 tienen pelo largo, además las imágenes 648, 759, 579 tienen el pelo claro.
4	<div> <div>label=Centroide 3</div> <div>x=1,119.548</div> <div>y=-329.6711</div> </div>	<div>Image 281</div>	<div>Image 3</div>	<div>Image 37</div>	<div>Image 301</div>	Se aprecia rasgos bastante jóvenes, la nariz a primera vista son de igual tamaño, los rostros están sonriendo, la imagen 301 sonríe levemente con respecto a las anteriores.

5		Image 268 	Image 44 	Image 49 	Image 660 	Entre las 4 imágenes se aprecia que todos tienen el pelo largo, y que sus narices tienen un tamaño bastante similar, se ven las imágenes bastante claras, no se aprecian sombras.
6		Image 443 	Image 711 	Image 771 	Image 292 	Entre las 4 imágenes se observan rostros muy finos y delgados, la gran mayoría no tiene expresiones fuertes, en la imagen 711, 771 y 292 se observa alguna sonrisa.
7		Image 27 	Image 186 	Image 114 	Image 395 	En si no se observa muchas características similares, aunque se pueden observar como los rostros están girados levemente totalmente independiente a la vista que es totalmente frontal.

3 Regresión (60%) .

Utilizar el dataset de la carpeta datos. 'Resultados_Saber_TyT_Gen_ricas_2020-1.csv' (ver [origen](#)), el caso de uso es que basado en las condiciones del estudiante vamos a predecir el puntaje que tendrá en las pruebas del saber. Por supuesto no se pueden utilizar ninguna variable de puntaje en las variables a utilizar o datos que se generen después de presentar el examen. La variable objetivo es MOD_INGLES_PUNT que muestra el nivel de inglés.

- Realizar la exploración de los datos correlación, scatter plots, boxplots e histogramas:

- ¿Qué variables son importantes para predecir el valor?

Las variables más importantes son:

ESTU_GENERO	object
ESTU_FECHANACIMIENTO	--> Tranformar object
ESTU_PAIS_RESIDE	object
ESTU_DISC_FISICA	object
ESTU_DEPTO_RESIDE	object
ESTU_MCPIO_RESIDE	--> Tranformar object
ESTU_AREARESIDE	object
ESTU_ESTADOCIVIL	object
ESTU_COLE_TERMINO	object
ESTU_PAGOMATRICULABECA	object
ESTU_PAGOMATRICULACREDITO	object
ESTU_PAGOMATRICULAPADRES	object

```

ESTU_PAGOMATRICULAPROPIO      object
ESTU_COMOCAPACITOEXAMENSB11    object
ESTU_SEMESTRECURSA             object
FAMI_EDUCACIONPADRE            object
FAMI_EDUCACIONMADRE            object
FAMI_TRABAJOLABORPADRE         object
FAMI_TRABAJOLABORMADRE         object
FAMI ESTRATOVIVIENDA           object
FAMI_TIENEINTERNET             object
FAMI_TIENESERVICIOTV           object
FAMI_TIENECOMPUTADOR           object
FAMI_TIENELAVADORA             object
FAMI_TIENEHORNOMICROOGAS       object
FAMI_TIENEAUTOMOVIL            object
FAMI_TIENEMOTOCICLETA          object
FAMI_TIENECONSOLAVIDEOJUEGOS  object
FAMI_CUANTOSCOMPARTEBAÑO       object
ESTU_VALORMATRICULAUNIVERSIDAD object
ESTU_HORASSEMANTRABAJA         object
INST_NOMBRE_INSTITUCION        object
GRUPOREFERENCIA               object
ESTU_PRGM_DEPARTAMENTO          object
ESTU_NIVEL_PRGM_ACADEMICO       object
ESTU_METODO_PRGM               object
ESTU_NUCLEO_PREGRADO           object
ESTU_INST_MUNICIPIO            object
ESTU_INST_DEPARTAMENTO          object
INST_CARACTER_ACADEMICO        object
INST_ORIGEN                    object
ESTU_PRIVADO_LIBERTAD          object
ESTU_ESTADOINVESTIGACION       object

```

1.2. Existen nulos?, ¿cómo se deben imputar?

Para observar si existen nulos, lo mejor sería crear una función en la que se pueda observar que variables tienen nulos, la mejor manera de imputar es por medio de la moda, lo más común aunque también se puede definir un valor específico, aunque para estos casos donde hay muchos valores lo mejor sería una transformación esto además de que elimina nulos acorta el tamaño cuando se va a realizar dummies variables.

En la celda 76 del notebook se verifican columnas del data frame nulas.

En la celda 78 y 79 del notebook se hacen transformaciones para limitar los datos e imputarlos.

En la celda 80 del notebook se vuelve a verificar que columnas presentan nulos.

En la celda 81 del notebook algunas columnas se imputan por medio de la moda y otras por relación a otras variables que ayudan a determinar la que se analiza.

En la celda 82 del notebook se elimina la columna de fecha de nacimiento debido a que se elaboró otra columna con la edad actual de la persona.

En la celda 83 del notebook se vuelve a verificar nulos y ya no presentan.

1.3. Crear dummy variables para incluirlas en la correlación

Para crear las dummies variables se tomarán los valores de tipo object, pues estos son los que se transformaran a dummies variables, después de identificarlos, se genera un dataframe de copia con

las columnas y en esa copia se generará los dummies, como los dummies se generan en enteros lo mejor es manejar al puntaje de inglés que es float para que no lo modifique.

En la celda 85 del notebook se observa los tipos de datos.

En la celda 86 del notebook se copia el dataframe para no modificar el original.

En la celda 90 del notebook se toman las columnas a las que se harán los dummies variables y al mismo tiempo se mantiene la del puntaje de inglés.

En la celda 91 del notebook se muestra el dataframe con los dummies.

- 1.4. Crear una correlación, que variables tienen un efecto positivo en el puntaje y cuales un efecto negativo.

Con respecto al puntaje de inglés, se observa que si tiene más recursos o su ubicación es de las mejores en oportunidades, puede llegar a determinar un mayor nivel de inglés.

Se puede decir que la relación entre recursos y educación puede llevar a un mayor nivel en el puntaje de inglés, también se debe tener en cuenta la situación de los padres debido a que resaltan en la educación del hijo aunque no es determinantes en algunos casos.

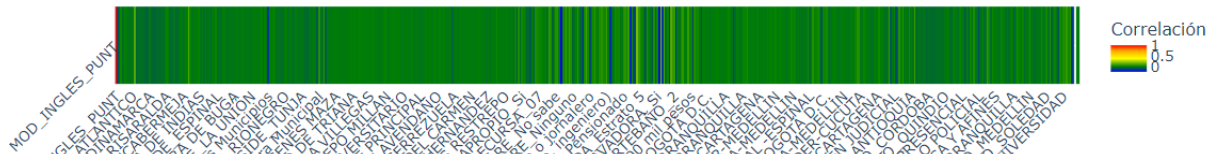
	MOD_INGLES_PUNT	ESTU_AGE	ESTU_GENERO_F	ESTU_GENERO_M	ESTU_PAIS_RESIDE_COLOMBIA	ESTU_PAIS_RESIDE_Otros Países	ES
MOD_INGLES_PUNT	1.000000	-0.075984	-0.031374	0.031374	-0.024311	0.024311	
ESTU_AGE	-0.075984	1.000000	-0.044470	0.044470	-0.007255	0.007255	
ESTU_GENERO_F	-0.031374	-0.044470	1.000000	-1.000000	0.004416	-0.004416	
ESTU_GENERO_M	0.031374	0.044470	-1.000000	1.000000	-0.004416	0.004416	
ESTU_PAIS_RESIDE_COLOMBIA	-0.024311	-0.007255	0.004416	-0.004416	1.000000	-1.000000	
...	
INST_ORIGEN_OFICIAL NACIONAL	-0.120117	-0.062737	-0.009391	0.009391	0.026377	-0.026377	
INST_ORIGEN_REGIMEN ESPECIAL	-0.021577	-0.028532	-0.075543	0.075543	-0.023333	0.023333	
ESTU_PRIVADO_LIBERTAD_N	NaN	NaN	NaN	NaN	NaN	NaN	
ESTU_ESTADOINVESTIGACION_PUBLICAR	0.011575	0.001742	-0.006010	0.006010	-0.002228	0.002228	
ESTU_ESTADOINVESTIGACION_VALIDEZ OFICINA JURÍDICA	-0.011575	-0.001742	0.006010	-0.006010	0.002228	-0.002228	

570 rows × 570 columns


```
Los 15 elementos más cercanos a 1:
MOD_INGLES_PUNT 1.000000
FAMI_TIENEHORNOMICROOGAS_Si 0.210109
FAMI_ESTRATOVIVIENDA_Estrato 3 0.184729
FAMI_TIENECONSOLAVIDEOJUEGOS_Si 0.138939
ESTU_DEPTO_RESIDE_BOGOTÁ 0.130574
ESTU_MCPIO_RESIDE_BOGOTÁ D.C. 0.130574
FAMI_TIENEAUTOMOVIL_Si 0.123851
FAMI_EDUCACIONMADRE_Técnica o tecnológica completa 0.112964
FAMI_EDUCACIONMADRE_Educación profesional completa 0.101637
FAMI_EDUCACIONPADRE_Educación profesional completa 0.100279
FAMI_EDUCACIONPADRE_Técnica o tecnológica completa 0.098920
ESTU_PRGM_DEPARTAMENTO_BOGOTÁ 0.097841
ESTU_NUCLEO_PREGRADO_INGENIERÍA DE SISTEMAS, TELEMÁTICA Y AFINES 0.093675
GRUPOREFERENCIA_TECNOLÓGICO EN ARTES - DISEÑO - COMUNICACIÓN 0.093192
FAMI_TRABAJOLABORPADRE_Trabaja como profesional (por ejemplo médico, abogado, ingeniero) 0.093001
Name: MOD_INGLES_PUNT, dtype: float64
```

```
Los 15 elementos más cercanos a -1:
FAMI_EDUCACIONPADRE_Primeria incompleta -0.232450
FAMI_TIENEHORNOMICROOGAS_No -0.210109
FAMI_TRABAJOLABORMADRE_Trabaja en el hogar, no trabaja o estudia -0.165853
FAMI_TRABAJOLABORPADRE_Trabaja por cuenta propia (por ejemplo plomero, electricista) -0.146510
FAMI_EDUCACIONMADRE_Secundaria (Bachillerato) completa -0.146066
FAMI_TIENECONSOLAVIDEOJUEGOS_No -0.138939
FAMI_CUANTOSCOMPARTEBANO_3 o 4 -0.127842
FAMI_TIENEAUTOMOVIL_No -0.123851
ESTU_MCPIO_RESIDE_Otros Municipios -0.122336
ESTU_VALORMATRICULAUNIVERSIDAD_No pago semestre -0.121363
INST_ORIGEN_OFICIAL NACIONAL -0.120117
FAMI_ESTRATOVIVIENDA_Estrato 2 -0.107595
FAMI_TRABAJOLABORPADRE_Es agricultor, pesquero o jornalero -0.096680
FAMI_ESTRATOVIVIENDA_Estrato 1 -0.092028
ESTU_NUCLEO_PREGRADO_FORMACIÓN RELACIONADA CON EL CAMPO MILITAR O POLICIAL -0.088830
Name: MOD_INGLES_PUNT, dtype: float64
```

Correlación con 'MOD_INGLES_PUNT'



En la celda 95 del notebook se realiza la correlación y se imprime para visualizar la matriz de correlación.

En la celda 96 del notebook se genera una función para tomar los valores más cercanos a 1 y a -1 para observar qué variables crecen proporcionalmente al inglés y que otras son inversamente proporcionales al puntaje de inglés.

En la celda 97 del notebook se genera un gráfico de la correlación para mirar dato por dato en un gráfico.

2. Divida los datos en training y testing

- 2.1. Aplique las transformaciones más importantes a los datos. (Hint calcular la edad basada en la fecha de nacimiento, agrupar variables categóricas con mucha cardinalidad en grupos).

Los datos ya se transformaron cuando se realizó la imputación y también cuando se realizan las dummies variables, pues estas generaban muchas columnas lo que conllevaba a que la correlación fuera muy lenta por ende se realizaron transformaciones con la finalidad de poder trabajar de una mejor manera.

En la celda 78 y 79 del notebook se hacen transformaciones para limitar los datos e imputarlos.

2.2. Entrenar un modelos de regresión

OLS Regression Results			
Dep. Variable:	MOD_INGLES_PUNT	R-squared:	0.247
Model:	OLS	Adj. R-squared:	0.240
Method:	Least Squares	F-statistic:	34.61
Date:	Sat, 23 Sep 2023	Prob (F-statistic):	0.00
Time:	12:15:09	Log-Likelihood:	-2.5323e+05
No. Observations:	54923	AIC:	5.075e+05
Df Residuals:	54407	BIC:	5.121e+05
Df Model:	515		
Covariance Type:	nonrobust		

En la celda 107 del notebook se empiezan a poner las variables que ya fueron transformadas en dummies en sus respectivos valores para la y será el puntaje de inglés, para las x será todas menos la del puntaje de inglés.

En la celda 88 del notebook se dividen los datos entre train y test, siendo test el 20% de los datos y train el 80%.

En la celda 109 del notebook se realiza la regresión y se muestra.

2.3. ¿Cuál es el mejor R squared?Cuál es el MAPE y el MSE.

[110]: #2.3 ¿Cuál es el mejor R squared?Cuál es el MAPE y el MSE.

```
[111]: #Para el train
predic_data_train = model_predict_value.predict(x_train)
r_squared_value_predict_train = r2_score(y_train, predic_data_train)
mape_value_abs_train = np.mean(np.abs((y_train - predic_data_train) / y_train)) * 100
mse_value_train = mean_squared_error(y_train, predic_data_train)
print("R^2:", r_squared_value_predict_train)
print("MAPE:", mape_value_abs_train)
print("MSE:", mse_value_train)
#MAPE, dió infinto debido a que y_train contiene ceros, por ende no es buena una medida adecuada de la precisión del modelo.

R^2: 0.24675977171337027
MAPE: inf
MSE: 591.9433604655856
```

```
[112]: #Para el test
predic_data_test = model_predict_value.predict(x_test)
r_squared_value_predict_test = r2_score(y_test, predic_data_test)
mape_value_abs_test = np.mean(np.abs((y_test - predic_data_test) / y_test)) * 100
mse_value_test = mean_squared_error(y_test, predic_data_test)
print("R^2:", r_squared_value_predict_test)
print("MAPE:", mape_value_abs_test)
print("MSE:", mse_value_test)
#MAPE, dió infinto debido a que y_test contiene ceros, por ende no es buena una medida adecuada de la precisión del modelo.

R^2: 0.23549679100809495
MAPE: inf
MSE: 597.1095809510937
```

Se puede observar que el R² para train es de 0.2467 muy parecido al R² del test es 0.2354, los mismo para MSE en train es 591.94 y en test es 597.1, se mantiene muy similares, en el caso de del MAPE para ambas es infinito debido que la y es 0 y genera infinito.

3. Remueva las variables que nos son relevantes

Para remover las variables que no son relevantes se observa por medio del P-value si este es menor que 0.005 significa que se conserva, es decir aporta a que la hipótesis nula no sea factible, en este caso se construye un código para determinar qué valores son menores que 0.005, pero esto no es una regla general que siempre se lleve a cabo en los modelos de regresión. En este caso se harán dos modelos en los cuales mantienen variables con un p-value menor a 0.005 y otro menor o igual a 0.005 y observa el comportamiento.

En la celda 120 del notebook se realiza una función para determinar las variables menores o iguales a 0.005, se realiza dos veces una con < 0.005 y otra con ≤ 0.005 , cabe resaltar que para p-value 0.005 hay menor variable que en ≤ 0.005 debido al porcentaje de datos que puede llegar a tomar debido al intervalo.

En la celda 136 del notebook se crea un dataframe vacío para cuando p-value es ≤ 0.005

En la celda 137 del notebook se realiza una función para eliminar todas las variables que el p-value no sea menor o igual que 0.005, es decir se obtienen las variables en las que el p-value es menor o igual que 0.005 y se analizan con las columnas que se tiene, luego se van comprando y si no las encuentra las elimina.

En la celda 138 del notebook se realiza lo mismo que la anterior exceptuando a que el p-value debe de ser menor a 0.005

En la celda 148 del notebook se realiza de nuevo el modelo de regresión con las variables que p-value ≤ 0.005 logrando obtener R^2 de 0.225

OLS Regression Results

Dep. Variable:	MOD_INGLES_PUNT	R-squared:	0.225
Model:	OLS	Adj. R-squared:	0.224
Method:	Least Squares	F-statistic:	145.0
Date:	Sat, 23 Sep 2023	Prob (F-statistic):	0.00
Time:	14:09:14	Log-Likelihood:	-2.5418e+05
No. Observations:	54923	AIC:	5.086e+05
Df Residuals:	54812	BIC:	5.096e+05
Df Model:	110		
Covariance Type:	nonrobust		

En la celda 151 del notebook se realiza de nuevo el modelo de regresión con las variables que p-value < 0.005 logrando obtener R^2 de 0.219

OLS Regression Results			
Dep. Variable:	MOD_INGLES_PUNT	R-squared:	0.219
Model:	OLS	Adj. R-squared:	0.218
Method:	Least Squares	F-statistic:	146.6
Date:	Sat, 23 Sep 2023	Prob (F-statistic):	0.00
Time:	14:09:27	Log-Likelihood:	-2.5405e+05
No. Observations:	54923	AIC:	5.083e+05
Df Residuals:	54817	BIC:	5.093e+05
Df Model:	105		
Covariance Type:	nonrobust		

4. Utilizando los datos de test medir el MAPE y el MSE de test. Qué tan diferentes son las métricas de training. (El menor error del grupo tiene un +1)

```
[ ]: #Análisis para P-value <= 0.005
```

```
[149]: #Para el test
predic_data_test = model_predict_value_delete.predict(x_test)
r_squared_value_predict_test = r2_score(y_test, predic_data_test)
mape_value_abs_test = np.mean(np.abs((y_test - predic_data_test) / y_test)) * 100
mse_value_test = mean_squared_error(y_test, predic_data_test)
print("R^2:", r_squared_value_predict_test)
print("MAPE:", mape_value_abs_test)
print("MSE:", mse_value_test)

R^2: 0.19612239395479136
MAPE: inf
MSE: 611.4988676080346

[150]: #Para el train
predic_data_train = model_predict_value_delete.predict(x_train)
r_squared_value_predict_train = r2_score(y_train, predic_data_train)
mape_value_abs_train = np.mean(np.abs((y_train - predic_data_train) / y_train)) * 100
mse_value_train = mean_squared_error(y_train, predic_data_train)
print("R^2:", r_squared_value_predict_train)
print("MAPE:", mape_value_abs_train)
print("MSE:", mse_value_train)

R^2: 0.2254089868204816
MAPE: inf
MSE: 612.708364323221
```

```
[ ]: #Análisis para P-value < 0.005
```

```
[152]: #Para el test
predic_data_test = model_predict_value_delete.predict(x_test)
r_squared_value_predict_test = r2_score(y_test, predic_data_test)
mape_value_abs_test = np.mean(np.abs((y_test - predic_data_test) / y_test)) * 100
mse_value_test = mean_squared_error(y_test, predic_data_test)
print("R^2:", r_squared_value_predict_test)
print("MAPE:", mape_value_abs_test)
print("MSE:", mse_value_test)
```

```
R^2: 0.21440536431959079
MAPE: inf
MSE: 628.3930985389898
```

```
[153]: #Para el train
predic_data_train = model_predict_value_delete.predict(x_train)
r_squared_value_predict_train = r2_score(y_train, predic_data_train)
mape_value_abs_train = np.mean(np.abs((y_train - predic_data_train) / y_train)) * 100
mse_value_train = mean_squared_error(y_train, predic_data_train)
print("R^2:", r_squared_value_predict_train)
print("MAPE:", mape_value_abs_train)
print("MSE:", mse_value_train)
```

```
R^2: 0.21929196643110316
MAPE: inf
MSE: 609.897685663433
```

En la celda 149 y 150 son para cuando p-value es menor o igual a 0.005

En la celda 152 y 153 son para cuando p-value es menor a 0.005

Para ambos casos si se observa el train y el test varían muy poco para el R^2 unos cuantos decimales, para el MSE un poco de valores y para el MAPE será infinito debido a la y que contiene ceros.

5. Describa en palabras que dice el modelo cuales son los principales hallazgos.

Cuando se realiza el modelo de regresión por primera vez de que el R^2 es de 0.247 es decir que el modelo logra explicar el 24.7%, que en realidad no es mucho, esto quiere decir el modelo no se ajusta bien a los datos y que la mayoría de la variabilidad de los datos usados no se está logrando capturar. El MAPE da infinito debido a valores que interpreta como cero, aun así si se colocara un datos muy cercano a cero esto forma un número muy demostrando así que si MAPE es muy grande tendiendo a infinito indica que el modelo no es adecuado para hacer pronósticos. Si se analiza el MSE al tener un valor tan grande como 591.94 confirma que el modelo está produciendo valores erróneos e imprecisiones.

Al quitar variables se observa que el modelo presenta mayores daños, se aprecia que R^2 disminuye significativamente y el valor de MSE se eleva dando a entender no es óptimo para predecir el puntaje de inglés, pero cuando se analizan los dos casos el p-value ≤ 0.005 y p-value < 0.005 , se parecía que cuando el p-value ≤ 0.05 el valor de R^2 no disminuye tanto, esto indica que la elección de variables es muy crucial, y es posible que se deban considerar variables con p-values más grandes para mejorar el modelo.