Tarea #: 2
Tema: Aprendizaje No supervisado y Regresión
Fecha entrega: 11:59 pm Septiembre 18 de 2023
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).

Nota: También se tiene la explicación detallada en los notebooks que se crearon, paso a paso para tener la información más organizada, se crearon 3 notebooks para no dejar en un archivo tanta información, repartiéndose de la siguiente manera, "Faces.ipynb" que contiene toda la información de las caras, "PruebasTyT.ipynb", es el primer acercamiento real que se tiene con los datos de las pruebas saber y el entrenamiento del primer modelo base, y "PruebasTyTMejorado.ipynb" donde contiene ahora el modelo con un análisis mas profundo con las variables numéricas y categóricas.

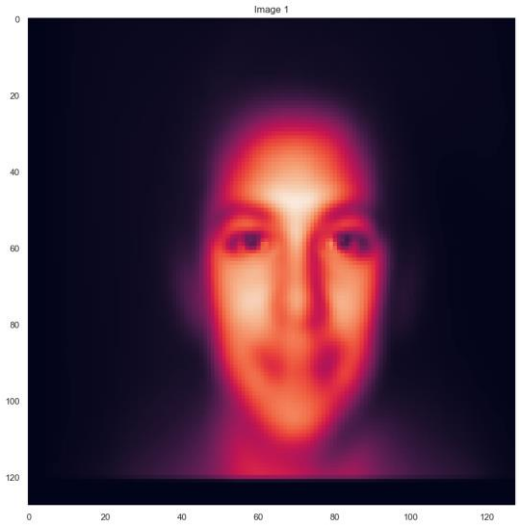
Contenido:

- PCA
- K-Means
- Regresión

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):

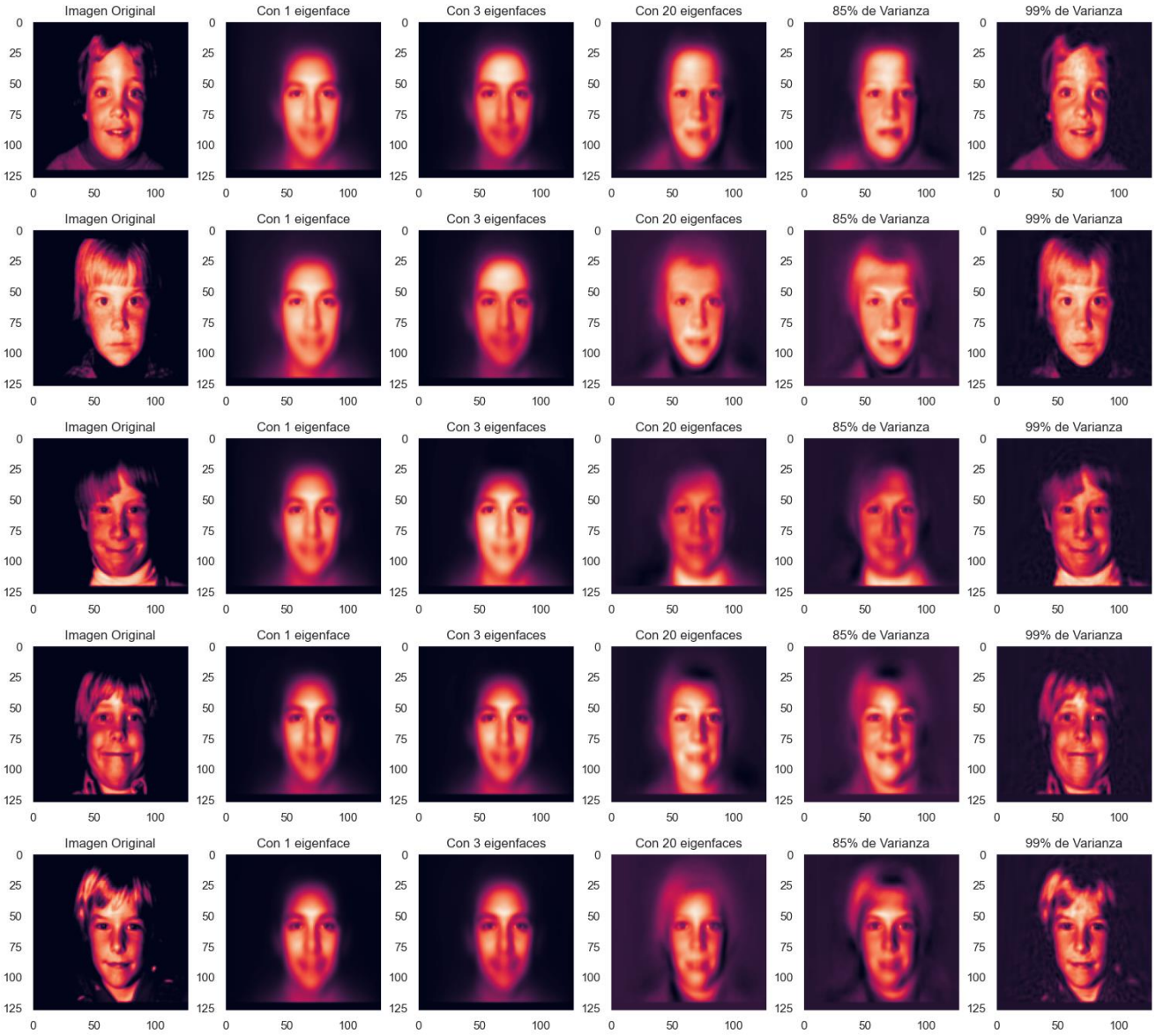
1. Calcular la mean face. Que es la cara con el promedio de los pixeles y visualizarla.



La imagen es sacada del repositorio en: Tarea 2/images/Mean_Face

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 10componentes, después con las componentes que explican el 90% de la varianzay por último con el número de componentes que tiene el 99% de la varianza. ¿Qué se puede concluir de los resultados?

- Para mantener el 90% de las características, se debe de conservar 75 componentes



Las imágenes son sacadas del repositorio Tarea 2/images/Face_Reconstruction

- Se concluye del análisis de las imágenes anteriores que:
  - Se debe de tener una gran cantidad de componentes para obtener el 99% de la varianza, es decir, entre mas porcentaje de varianza se requiera, aumentara casi que exponencialmente los componentes requeridos
  - Por otro lado, entre mas componentes se tenga la imagen va a tener mucha más definición o nitidez

- Del mismo modo entre menos componentes se tengan mas se parecerá a la cara media, pero es por el hecho de que no existen muchos componentes por los cuales reconstruir, y como en la formula el valor medio esta presente, entonces ese es el que va a predominar si existen pocas componentes.

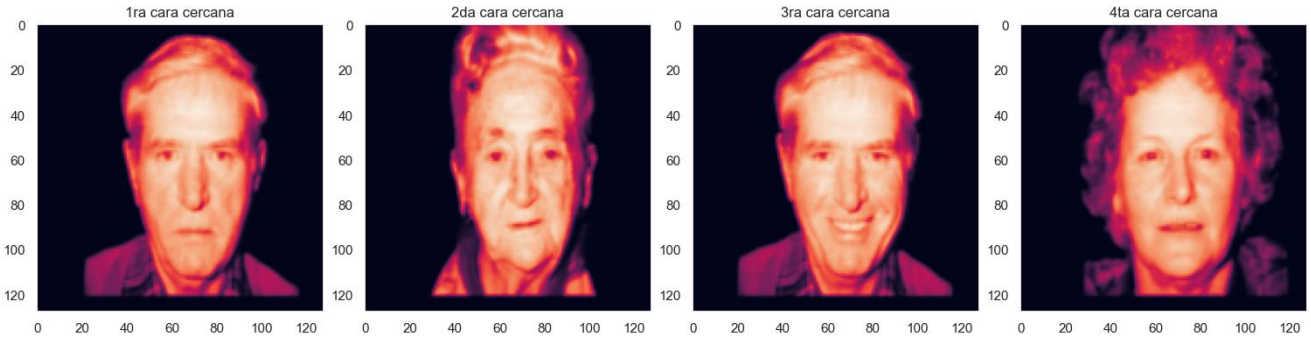
K-means (20%)

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(predecir con los centroides el cluster de un nuevo dato).

- La creación del k-means con su explicación está en el repositorio

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

Cluster 1:



Se evidencia que las personas en la imagen son personas de la tercera edad.

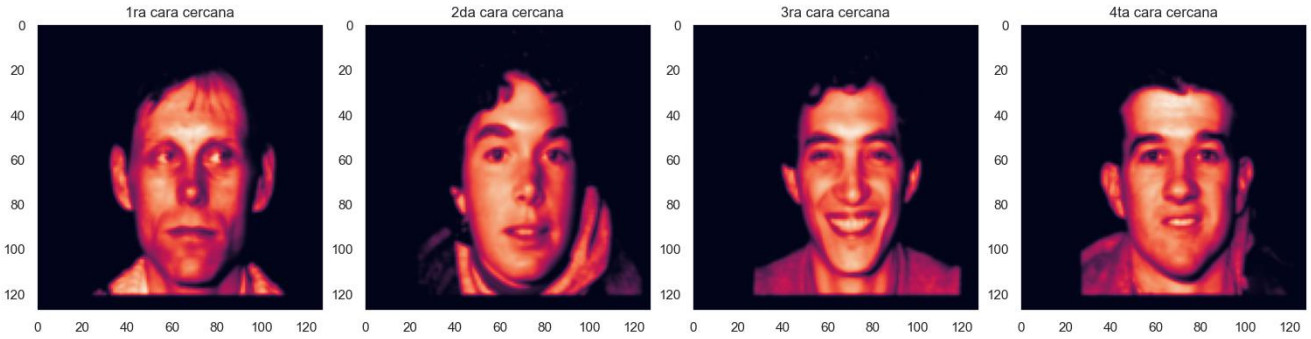
Cluster 2:

Se evidencia que son personas con cara seria y además con el pelo medio largo.

Cluster 3:

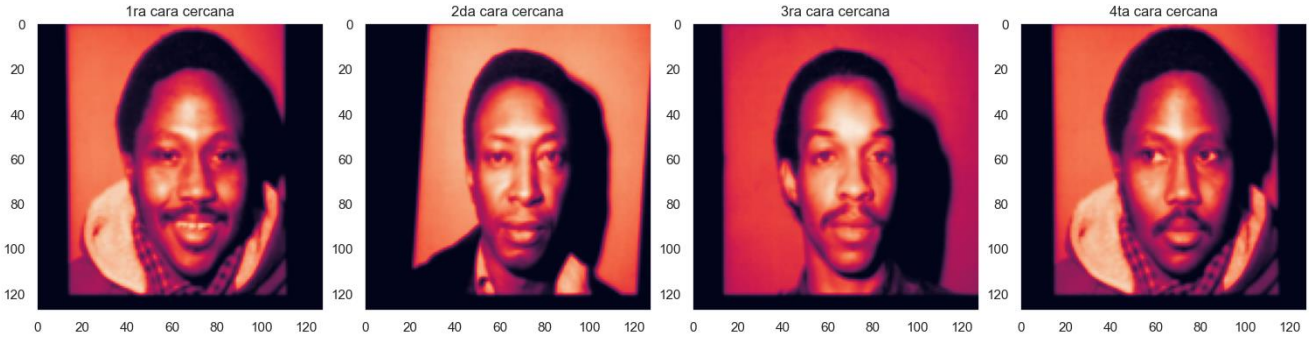
Se evidencia que dos de las cuatro imágenes son del mismo sujeto, además de que la mayoría tiene le pelo rizado.

Cluster 4:



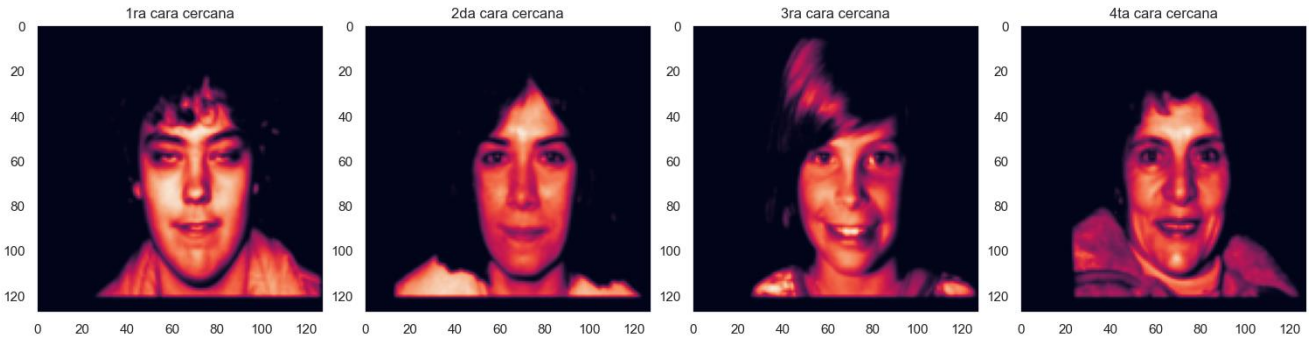
No se evidencia claramente que relación tienen en común, aunque podría decir a opinión que tienes facciones un poco parecidas entre sí.

Cluster 5:



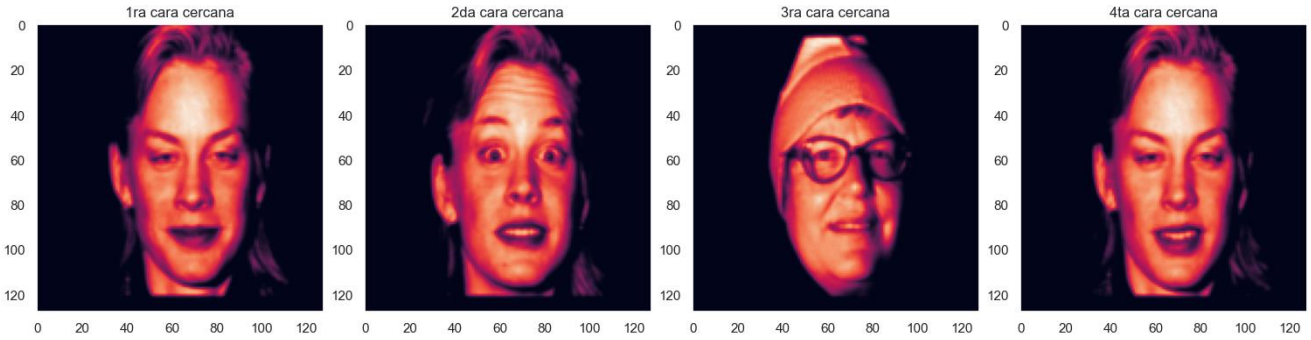
Se evidencia que son personas de color.

Cluter 6:



Son personas que tienen una expresión de felicidad o alegría en su rostro.

Cluster 7:



Se evidencia que se relacionaron tres imágenes de la misma persona con una expresión diferente en su rostro, existe otra imagen que no logro encontrar algo en común.

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.

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

2.1. ¿Qué variables son importantes para predecir el valor?

Según el resultado obtenido en el entramiento del último modelo, las variables que tienen un error bajo y un p-value bajo son:

['PUNT_GLOBAL', 'EDAD', 'MCPIO_VIVE_ESTUDIA',
'DUMMY_ESTU_ESTADOCIVIL_Soltero',
'DUMMY_ESTU_PAGOMATRICULAPADRES_Si',
'DUMMY_ESTU_COMOCAPACITOEXAMENSB11_Repasó por cuenta propia',
'DUMMY_FAMI_TIENESERVICIOTV_Si',
'DUMMY_FAMI_TIENECOMPUTADOR_Si',
'DUMMY_FAMI_TIENELAVADORA_Si',
'DUMMY_FAMI_TIENEHORNOMICROOGAS_Si',
'DUMMY_FAMI_TIENEAUTOMOVIL_Si',
'DUMMY_FAMI_TIENEMOTOCICLETA_Si',
'DUMMY_ESTU_METODO_PRGM_PRESENCIAL',
'DUMMY_GROUP_ESTU_DEPTO_RESIDE_Grupo 2',
'DUMMY_GROUP_ESTU_MCPIO_RESIDE_Grupo 4',
'DUMMY_GROUP_FAMI_EDUCACIONPADRE_Grupo 4',
'DUMMY_GROUP_FAMI_EDUCACIONMADRE_Grupo 4',
'DUMMY_GROUP_FAMI_TRABAJOLABORPADRE_Grupo 3',
'DUMMY_GROUP_FAMI_TRABAJOLABORPADRE_Grupo 4',
'DUMMY_GROUP_FAMI_TRABAJOLABORMADRE_Grupo 4',
'DUMMY_GROUP_FAMI_ESTRATOVIVIENDA_Grupo 4',
'DUMMY_GROUP_FAMI_CUANTOSCOMPARTEBAÑO_Grupo 4',
'DUMMY_GROUP_ESTU_VALORMATRICULAUNIVERSIDAD_Grupo 4',
'DUMMY_GROUP_ESTU_HORASSEMANATRABAJA_Grupo 4',
'DUMMY_GROUP_ESTU_PRGM_ACADEMICO_Grupo 4',
'DUMMY_GROUP_INST_CARACTER_ACADEMICO_Grupo 4',
'DUMMY_GROUP_INST_ORIGEN_Grupo 2',
'DUMMY_GROUP_INST_ORIGEN_Grupo 4']

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

En un primer acercamiento, se imputaron de forma muy básica, ya que a los numéricos se les puso un 0 y a los categóricos se les creo una categoría denominada "SIN_CATEGORIA". Pero después de hacer el primer modelo base y un análisis mas

concreto, se opto por imputar con la moda para datos categóricos y con la media para datos numéricos

2.3. Crear dummy variables para incluirlas en la correlación
2.4. Crear una correlación, que variables tienen un efecto positivo en el puntaje y cuales un efecto negativo.

Para este caso, me resulto un escenario en el cual no existía una correlación negativa que afectara significativa el modelo, por ende, no agregaré ninguna negativa, pero de las positivas obtuve las siguientes:

['MOD_LECTURA_CRITICA_PUNT',
'MOD_LECTURA_CRITICA_PNAL',
'MOD_LECTURA_CRITICA_PNBC',
'MOD_COMPETEN_CIUDADA_PUNT',
'MOD_COMPETEN_CIUDADA_PNAL',
'MOD_COMPETEN_CIUDADA_PNBC',
'MOD_INGLES_PNAL',
'MOD_INGLES_PNBC',
'PUNT_GLOBAL',
'PERCENTIL_GLOBAL',
'PERCENTIL_NBC']

3. Divida los datos en training y testing
3.1. 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).

En el caso de la fecha de nacimiento, se hizo la conversión a la edad, pero a la final, en la matriz de correlación no tuvo un impacto significativo en la variable de estudio, para las categorías como por ejemplo el departamento, municipio, que estudios tenían los papas, etc. Se agrupaban de a cuatro grupos ayudándonos con los cuartiles, para que los grupos estuvieran distribuidos de manera equitativa.

3.2. Entrenar un modelo de regresión
3.3. ¿Cuál es el mejor R squared? Cuál es el MAPE y el MSE.

- Luego de crear el modelo base, se obtuvo un **R-squared de 0.909**, pero la explicación es bastante evidente, lo que ocurrió era que lo había entrenado con todos los datos (excepto la variable de estudio) y al entrenarse de esta forma pues existían muchos datos con una alta colinealidad, lo que afectaba mucho al modelo, además de que

también existían variables con una alta correlación entre si lo que también afectaba el resultado del modelo, por ende, aunque se observara un buen R-square, era porque realmente el modelo estaba mal entrenado,

- Por ende el siguiente paso era entrenarlo pero ahora con los valores que tuvieran una impacto en la correlación significativo para incluirlos (es decir menor a -0.4 y mayor a 0,4), después de agregar las variables que tienen un impacto con la correlación significativo, el modelo obtuvo un **R-square del 0.909**, seguía siendo un buen R-square, pero el modelo seguía entrenándose mala forma, ya que aunque solo se estaba entrenando con datos que tuvieran buena correlación , pero sin tener en cuenta que también existía una alta correlación entre ellos mismos, por ende se debía de descartar los datos que tuvieran alta correlación entre si ya que esto también afectaba negativamente al modelo.
- En el siguiente entrenamiento solo de dejaron las variables que no tenían una alta correlación entre sí, y luego al entrenar el modelo se obtuvo un **R-square de 0.577**, pero se observaba que los valores nulos tenían un gran impacto en el modelo ya que las únicas variables categóricas que afectaban al modelo eran los datos nulos que estaban en la categoría 'SIN_CATEGORIA' así que se tomó la decisión de imputarlo con la mediana y moda como se dijo en un punto anterior.
- En el próximo entrenamiento se imputaron las variables categóricas de mejor forma y se volvió a entrenar el modelo con todas las variables categóricas, esta vez solo se iban a dejar las que tenían un p-value menor a 0.005, en ese caso se tuvo un **R-square de 0.535** aquí el resultado tenía más sentida ya que los valores estaban imputados de mejor forma, para este caso te tenia, error calculado en train **MSE: 364.22926763921305**, **MSE test: 368.7153893244717**, es decir una desviación del valor real aproximada de 19 puntos. Cabe aclarar que el MAPE no se podía sacar ya que como existían valores igual a 0 en los puntajes no se podía hacer la división ya que el resultado era infinito.
- Por ultimo se noto que seguían resultando p-values bastante altos, así que se volvió a recortar las variables necesarias para el entrenamiento, en este se eliminaron **28 variables** que tenían un p-value muy alto, se volvió a entrenar le modelo y esta vez se obtuvo un **R-Square de 0.531**, es decir, se logro mantener el porcentaje restando una cantidad significativa de variables, por ende se concluye que se realizó un buen trabajo, en este caso el **MSE para train fue de 365.38206101824244** y **MSE para test fue de 368.8696995837154**, es decir aún se conserva la desviación del valor original aproximada de 19 puntos.

4. Remueva las variables que nos son relevantes
5. 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)

   Como se menciono en el punto anterior, según el mejor resultado, la diferencia entre el MSE de train y test es de 3 puntos aproximadamente, es decir no existe una diferencia significativa por ende se puede concluir que el modelo se entreno de forma correcta evitando overfitting u otros procesos que entorpecieran el entrenamiento.

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

   - Se evidencia que entre mas variables existan con una alta correlación entre sí, puede afectar el modelo, aunque se obtenga un R-square mucho mejor, realmente está empeorando el modelo.
   - Aunque se reduzcan muchas variables en el modelo se puede conservar un buen porcentaje de R-quare y MSE, casi sin cambiar sus valores, por ende, entre más datos no significa que sea mejor el modelo, y mucho menos que sea eficiente.
   - También se evidencia que los datos nulos tenían un gran impacto a la hora de predecir el puntaje del inglés, por ende, se deben imputar de forma adecuada.
   - El modelo final puede predecir de forma casi precisa los resultados, solo con la posibilidad de desviarse 19 puntos del valor original, que, aunque puede ser mucho, realmente considero que fue un buen trabajo para ser la primer vez en entrar en contacto con la construcción de un modelo.