**Curso:** Estadistica 3
**Profesor:** Jose Daniel Ramirez Soto

**Tarea** #: 2
**Tema**: Regresión
**Fecha entrega**: 11:59 pm octubre 16 de 2024
**Objetivo**: Aplicar los conceptos de KNN, regresión y GBM 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).

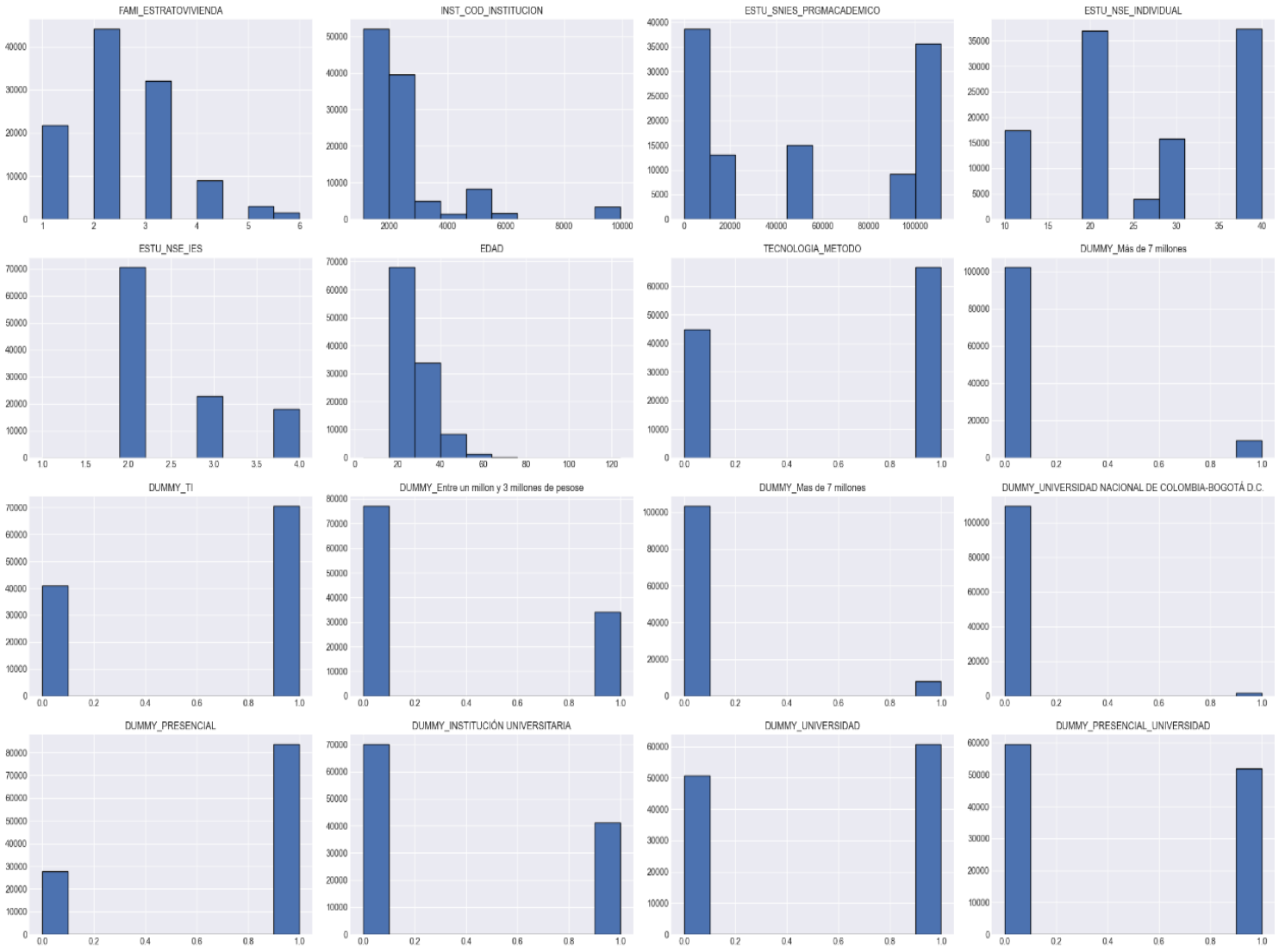
# 1 Regresión (60%)

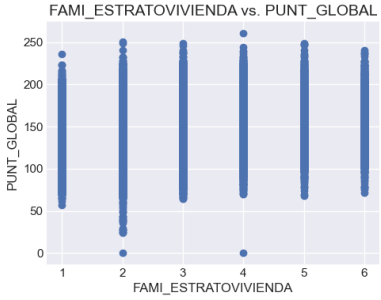
Utilizar kaggle para descargar la base https://www.kaggle.com/t/efa882e3a6d94bf799278c56ef3c8317 , el caso de uso es que basado en las condiciones del estudiantes vamos a predecir el puntaje que tendrá en las pruebas del saber en lecture critica “PUNT_GLOBAL”.

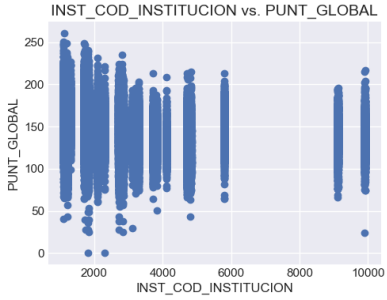
1. Realizar la exploración de los datos correlación, scatter plots, boxplots e histogramas:

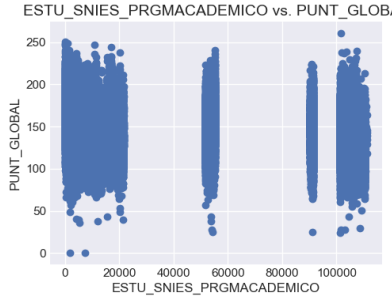
- boxplot de algunas variables

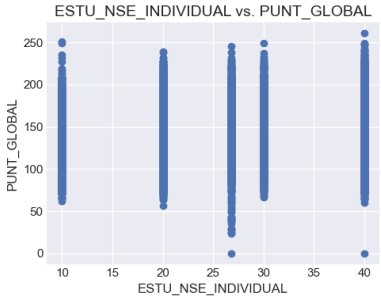
- Histogramas algunas variables

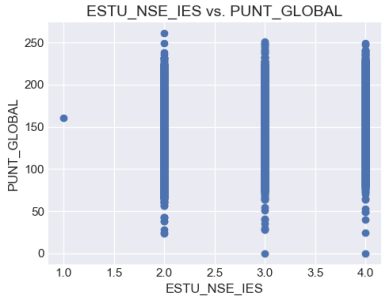


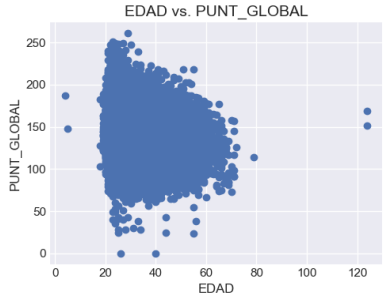

FAMI_ESTRATOVIVIENDA vs. PUNT_GLOBAL
PUNT_GLOBAL
FAMI_ESTRATOVIVIENDA

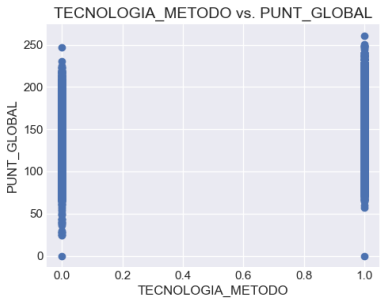

INST_COD_INSTITUCION vs. PUNT_GLOBAL
PUNT_GLOBAL
INST_COD_INSTITUCION

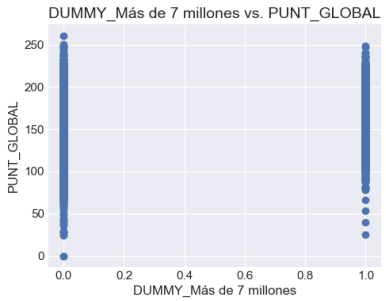

ESTU_SNIES_PRGMACADEMICO vs. PUNT_GLOBAL
PUNT_GLOBAL
ESTU_SNIES_PRGMACADEMICO

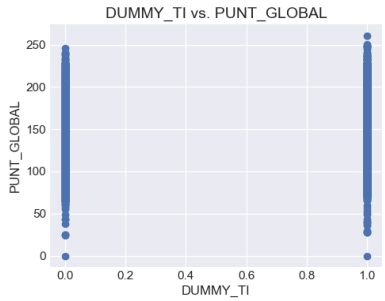

ESTU_NSE_INDIVIDUAL vs. PUNT_GLOBAL
PUNT_GLOBAL
ESTU_NSE_INDIVIDUAL

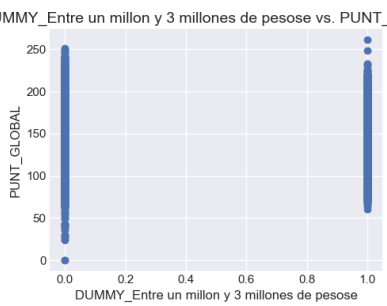

ESTU_NSE_IES vs. PUNT_GLOBAL
PUNT_GLOBAL
ESTU_NSE_IES

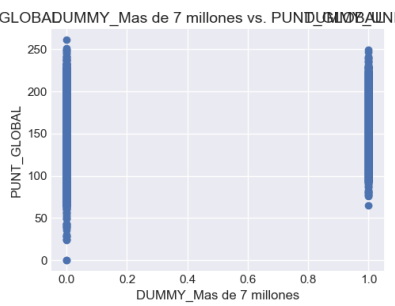

EDAD vs. PUNT_GLOBAL
PUNT_GLOBAL
EDAD

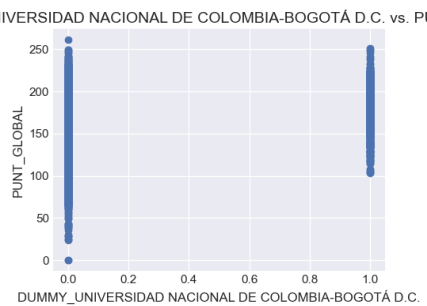

TECNOLOGIA_METODO vs. PUNT_GLOBAL
PUNT_GLOBAL
TECNOLOGIA_METODO

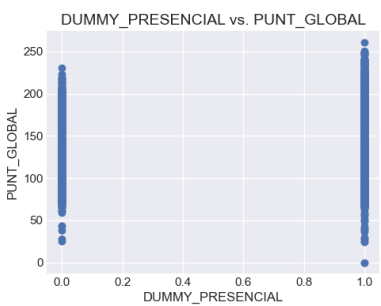

DUMMY_Más de 7 millones vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_Más de 7 millones

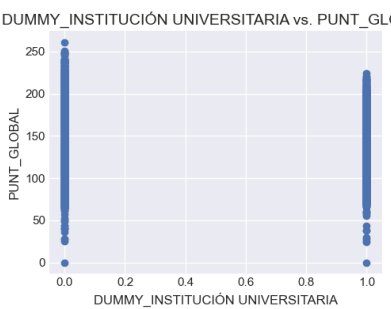

DUMMY_TI vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_TI

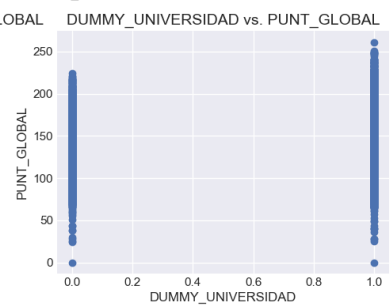

DUMMY_Entre un millon y 3 millones de pesose vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_Entre un millon y 3 millones de pesose

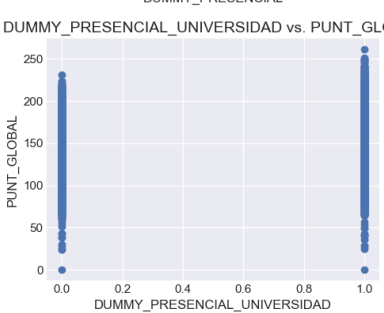

DUMMY_Mas de 7 millones vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_Mas de 7 millones


DUMMY_UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C. vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C.


DUMMY_PRESENCIAL vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_PRESENCIAL


DUMMY_INSTITUCIÓN UNIVERSITARIA vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_INSTITUCIÓN UNIVERSITARIA


DUMMY_UNIVERSIDAD vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_UNIVERSIDAD


DUMMY_PRESENCIAL_UNIVERSIDAD vs. PUNT_GLOBAL
PUNT_GLOBAL
DUMMY_PRESENCIAL_UNIVERSIDAD

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

- ['FAMI_ESTRATOVIVIENDA',
- 'INST_COD_INSTITUCION',
- 'ESTU_SNIES_PRGMACADEMICO',
- 'ESTU_NSE_INDIVIDUAL',
- 'ESTU_NSE_IES',
- 'EDAD',
- 'TECNOLOGIA_METODO',
- 'DUMMY_Más de 7 millones',
- 'DUMMY_TI',
- 'DUMMY_Entre un millon y 3 millones de pesose',
- 'DUMMY_Mas de 7 millones',
- 'DUMMY_UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C.',
- 'DUMMY_PRESENCIAL',
- 'DUMMY_INSTITUCIÓN UNIVERSITARIA',
- 'DUMMY_UNIVERSIDAD',
- 'DUMMY_PRESENCIAL_UNIVERSIDAD']

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

Se aplicaron dos formas:

- Si existen mas del 50% de datos nulos la columna será eliminada
- Las columnas con nulos que no fueron eliminadas se llenaran los datos con la media.

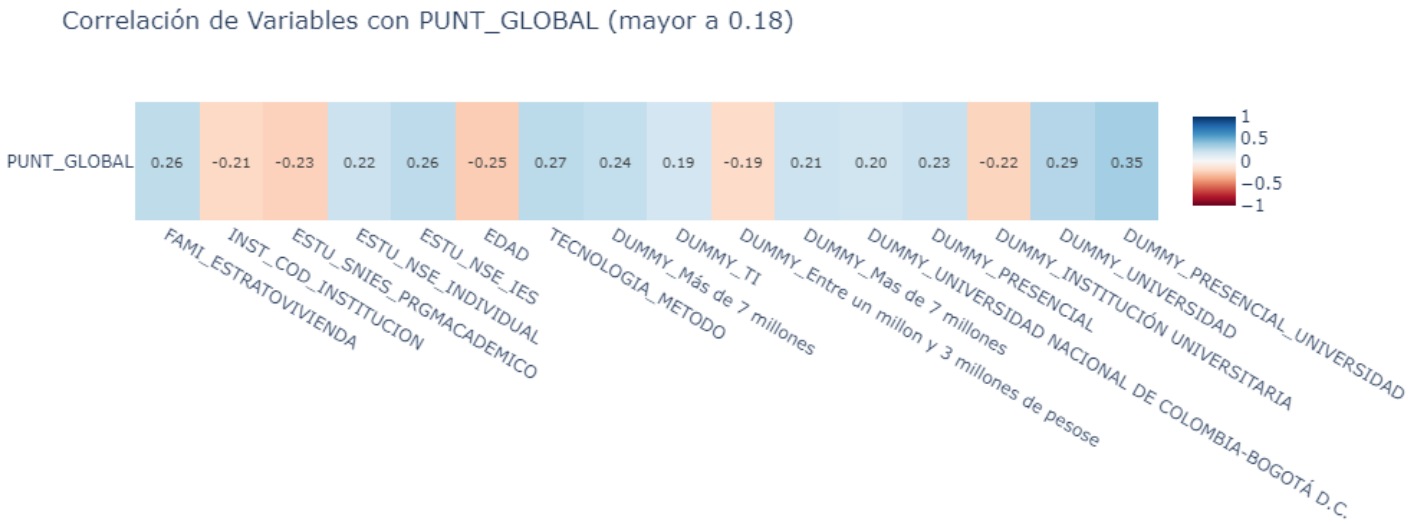
```
# umbral para eliminar columnas con demasiados valores nulos
threshold = 0.5  # 50% de nulos

initial_size = df.shape

columns_to_remove = df.columns[df.isnull().mean() > threshold]


for column in columns_to_remove:
    df.drop(column, axis=1, inplace=True)


# Llenar los valores nulos en columnas numéricas con la media de cada columna
numeric_cols = df.select_dtypes(include=np.number).columns
for col in numeric_cols:
    mean_value = df[col].mean()
    df[col].fillna(mean_value, inplace=True)
    print(f"Llenada columna '{col}' con la media: {mean_value}")
```

1.3. Crear dummy variables para incluirlas en la correlación

- En la siguiente imagen se pueden ver las dummy variables que se incluyeron

en la correlación

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




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).
   2.2. Entrenar un modelo de regresión
   2.3. ¿Cuál es el mejor R squared? Cuál es el MAPE y el MSE.

```
Métricas para el conjunto de entrenamiento:
MSE: 381.5183
R2: 0.3055

Métricas para el conjunto de prueba:
MSE: 387.7251
R2: 0.2999
```

3. Remueva las variables que nos son relevantes

   - Todas las variables más relevantes para el modelo están en la variable filtered_corr_df['Variable']

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)

```
Métricas para el conjunto de entrenamiento:
MSE: 381.5183
R2: 0.3055

Métricas para el conjunto de prueba:
MSE: 387.7251
R2: 0.2999
```

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

    - El modelo de regresión lineal muestra que, al agregar más variables, su rendimiento mejora un poco. Para optimizarlo, se utilizó un *Pipeline* que normaliza los datos para que todas las variables tengan el mismo peso, transforma las características en polinomios de grado tres para captar relaciones no lineales, y aplica un modelo de regresión *Ridge* con un parámetro de regularización de 10. Este enfoque ayuda a que el modelo no se ajuste demasiado a los datos y, gracias a estas técnicas, el modelo puede mejorar un poco mas.

# 2 Crear un modelo de KNN (20%)

Utilizar los datos para crear un modelo de KNN que permita predecir el puntaje por estudiante. Utilizar kaggle para descargar la base https://www.kaggle.com/t/efa882e3a6d94bf799278c56ef3c8317 y las mismas transformaciones dle punto anterior

1) Hacer pruebas con 5, 10, 20 y 30 vecinos. Seleccione el numero de vecinos basado en el error de test MSE.

| vecinos | MSE train | MAPE train | MSE test | MAPE test |
|---|---|---|---|---|
| 5 | 270.5349 | 6264542243941.8223 | 410.2575 | 33101032888956.7383 |
| 10 | 311.0558 | 6754591113024.3672 | 383.3319 | 29564597751247.6914 |
| 20 | 337.4176 | 7153703078565.8242 | 376.9832 | 28645124615443.3398 |
| 30 | 351.2488 | 7061082158292.4932 | 378.5456 | 28048982692229.5312 |

2) Describa cual es mejor modelo entre la regresion o el knn.

- El mejor modelo con 16 variables que se utilizaron fue el knn con K = 20, con un MAPE en test de 376.9832 en cambio la regresión obtuvo un MAPE de 387.7251 se puede ver que la diferencia es un poco grande, siendo así el KNN un poco mejor a la hora de predecir los datos que la regresion

# 3 Crear un modelo de GBM (20%)

Entrenar un modelo de GBM y hacer la prediccion. Cual es el MSE y MAPE para train y test.

```
Train MSE: 281.0551
Train R²: 0.4883
Train MAPE: 608670589913726.50%
Test MSE: 354.4139
Test R²: 0.3600
Test MAPE: 2749280772630044.00%
```