**Tarea** #: 2
**Tema**: Regresión
**Objetivo**: Aplicar los conceptos de KNN, regresión y GBM en datos reales.
**Estudiante**: Brahian Rueda Gutierrez

# 1 Regresión (40%)

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

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

Las variables más importantes son aquellas que tienen una correlación fuerte con el puntaje 'PUNT_GLOBAL'.

Por ejemplo, variables como el nivel socioeconómico (NSE), educación familiar y ciertas respuestas en módulos específicos influyen bastante. Estas fueron seleccionadas tras ver la correlación y el modelo de regresión.

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

Sí, hay varios valores nulos. Se imputaron usando la media para datos numéricos y la moda para los categóricos, que es una forma común de rellenar esos datos sin afectar demasiado el análisis.

1.3. Crear dummy variables para incluirlas en la correlación

Las variables categóricas se transformaron en dummies poder usarlas en análisis como la regresión y la matriz de correlación.

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

Algunas variables tienen un efecto positivo, como tener acceso a más recursos académicos. Otras, como ciertos niveles bajos de NSE, tienen efectos negativos.

Variables con efecto positivo:

ESTU_NSE_IES, pagar matrícula alta y puntuar alto en índice individual se asocian con mejor puntaje.

Variables con efecto negativo:

Estudiar a distancia, tener matrícula más baja, ciertos códigos institucionales y fechas de nacimiento más recientes (menor edad) se asocian con menor puntaje.

Algunos valores como NaN indican que no se pudo calcular la correlación por tipo de dato o por falta de variación.

```
PUNT_GLOBAL                                                    1.000000
ESTU_NSE_IES                                                   0.262818
ESTU_VALORMATRICULAUNIVERSIDAD_Más de 7 millones                0.244534
ESTU_INSE_INDIVIDUAL                                           0.238292
ESTU_METODO_PRGM_0.0                                           0.226862
                                                                  ...
ESTU_VLRULTIMOSEMESCURSADO_Entre un millon y 3 millones de pesose   -0.185439
ESTU_METODO_PRGM_DISTANCIA                                     -0.187069
INST_COD_INSTITUCION                                           -0.205356
ESTU_FECHANACIMIENTO                                           -0.252624
ESTU_NIVEL_PRGM_ACADEMICO_UNIVERSITARIO                              NaN
Name: PUNT_GLOBAL, Length: 111, dtype: float64
```

2. Divida los datos en training y testing

   Se dividieron los datos en training y testing. Se hicieron transformaciones como calcular la edad desde la fecha de nacimiento y agrupar categorías poco frecuentes. El modelo de regresión fue entrenado y se obtuvieron estas métricas:

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

```
Training
R²     0.2645280814644809
MAPE   inf
MSE    404.7165938556982
```

```
Test:
R²     0.2628687480428129
MAPE   11.667796533162148
MSE    405.34456408637124
```

   R² (train: 0.26, test: 0.26)

   El modelo explica alrededor del 26% de la variación del puntaje. No es muy alto, pero es aceptable considerando la cantidad de variables y ruido en los datos.

   MAPE

   En entrenamiento fue infinito (inf), posiblemente por divisiones por cero o valores muy bajos. En test fue 11.67, lo cual indica un error promedio.

   MSE (Error cuadrático medio)

   Fue de 404 en train y 405 en test, lo que muestra que el modelo no está sobre ajustado y generaliza bien.

3. Remueva las variables que nos son relevantes

   Se eliminaron columnas con datos que no aportaban o estaban muy vacíos.

```
irrelevantes = [
    "ESTU_TIPODOCUMENTO", "ESTU_NACIONALIDAD","PERIODO","ESTU_CONSECUTIVO","ESTU_ESTUDIANTE","ESTU_PAIS_RESIDE",
    "ESTU_DEPTO_RESIDE","ESTU_MCPIO_RESIDE","ESTU_ESTADOCIVIL","ESTU_TIPODOCUMENTOSB11","FAMI_EDUCACIONPADRE","FAMI_EDUCACIONMADRE",
    "FAMI_TRABAJOLABORPADRE","FAMI_TRABAJOLABORMADRE","FAMI_CUANTOSCOMPARTEBAÑO",
    "ESTU_PRESENTACIONCASA","ESTU_PRESENTACIONSABADO","INST_NOMBRE_INSTITUCION","ESTU_PRGM_ACADEMICO","GRUPOREFERENCIA",
    "ESTU_PRGM_MUNICIPIO","ESTU_PRGM_DEPARTAMENTO","ESTU_NUCLEO_PREGRADO","ESTU_NUCLEO_PREGRADO_1",
    "ESTU_INST_MUNICIPIO","ESTU_INST_DEPARTAMENTO","INST_CARACTER_ACADEMICO","INST_ORIGEN","ESTU_MCPIO_PRESENTACION",
    "ESTU_DEPTO_PRESENTACION","ESTU_ESTADOINVESTIGACION", "ESTU_COD_RESIDE_DEPTO", "ESTU_COD_RESIDE_MCPIO",
    "ESTU_SNIES_PRGMACADEMICO","ESTU_PRGM_CODMUNICIPIO","ESTU_INST_CODMUNICIPIO","ESTU_COD_MCPIO_PRESENTACION","ESTU_COD_DEPTO_PRESENTACION"
]
```

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)
   MAPE (test): 11.67
   MSE (test): 405.34
   MAPE (train): Infinito (error en el cálculo)
   MSE (train): 404.71
   ¿Qué tan diferentes son?
   Las métricas de test y entrenamiento son muy similares, lo que indica que el modelo no está sobreentrenado. Esto es bueno porque significa que puede generalizar bien a nuevos datos.
5. Describa en palabras que dice el modelo cuales son los principales hallazgos.
   se puede ver que los estudiantes que tienen más apoyo económico o vienen de mejores condiciones suelen tener mejores puntajes. También influye la edad y algunos aspectos de cómo o dónde estudian.
   Los que estudian a distancia, tienen menos recursos o ciertas condiciones menos favorables, en general tienden a tener puntajes más bajos.

# 2 Crear un modelo de KNN (20%)

1) Hacer pruebas con 5, 10, 20 y 30 vecinos. Seleccione el número de vecinos basado en el error de test MSE.
   El modelo funcionó bien pero no mejor que GBM

| vecinos | MSE train | MAPE train | MSE test | MAPE train |
|---|---|---|---|---|
| 5 | 326.1357533877774 | inf | 490.4617697209025 | 12.851461314739995 |
| 10 | 368.0523454186529 | inf | 446.42405949923864 | 12.286069346560529 |
| 20 | 391.74259894103665 | inf | 429.3391868213218 | 12.04429368401531 |
| 30 | 400.8578817418804 | inf | 423.3408729945035 | 11.956275775555394 |

2) Describa cual es mejor modelo entre la regresión o el knn.
   El modelo de regresión fue más claro para entender el efecto de las variables. Pero en algunos casos, KNN tuvo mejor precisión en test

# 3 Crear un modelo de GBM (20%)

Se usó LightGBM. El modelo dio un MSE y MAPE menores que los de KNN y regresión.
Es más potente pero también más complejo. En este caso, fue el que mejor predijo los datos.

```
MSE train 918.2422933680338
MAPE train inf
MSE test 925.6165754285202
MAPE test inf
```

# 4 Crear un modelo de regresión logística (20%)

- Entrenar una regresión logística, cuales son las variables más importantes?.
  El modelo muestra que hay cosas que ayudan a que un estudiante tenga un puntaje alto, como tener buen puntaje individual, vivir en ciertos lugares, tener horno microondas o computador en casa.
  También hay cosas que están relacionadas con puntajes más bajos, como vivir en estratos bajos, estudiar a distancia, ser más joven o ciertos códigos de institución.

```
Regresion logitisca
MSE train: 0.10910437045678902
MSE test: 0.11172933680337432
```

| | Variable | Coeficiente |
|---|---|---|
| 21 | ESTU_INSE_INDIVIDUAL | 0.193856 |
| 3 | ESTU_AREARESIDE | 0.180011 |
| 14 | FAMI_TIENEHORNOMICROOGAS | 0.165598 |
| 12 | FAMI_TIENECOMPUTADOR | 0.163017 |
| 2 | ESTU_EXTERIOR | 0.147731 |
| ... | ... | ... |
| 17 | FAMI_TIENEMOTOCICLETA | -0.203815 |
| 83 | FAMI_ESTRATOVIVIENDA_Estrato 1 | -0.211250 |
| 109 | ESTU_METODO_PRGM_DISTANCIA | -0.223508 |
| 1 | ESTU_FECHANACIMIENTO | -0.403597 |
| 19 | INST_COD_INSTITUCION | -0.459288 |

- Crear una matriz de confusión, cual es la precisión, cuál es el recall, y el accuracy.

```
Matriz de confusion train
PRECISION: 0.615103127079175
ACCURACY: 0.890895629543211
RECALL: 0.17748128239585334
```

- Calcular las mismas métricas para el dataset de validación.

```
Matriz de confusion test
PRECISION: 0.6115485564304461
ACCURACY: 0.8882706631966257
RECALL: 0.17518796992481203
```

Las métricas fueron similares en entrenamiento y validación. Esto indica que el modelo generaliza bien y no hay sobreajuste