# Sustentación sobre Tarea 2

## Exploración de datos

El proceso llevado para realizar los modelos fue explorar de forma visual las variables que componían el dataset y que significaba cada uno.

Después de esto se seleccionó las variables importantes descartando variables en donde no había variación en los datos, había muchos nulos o no era significativos para predecir el PUNT\_GLOBAL.

- ESTU\_GENERO

- ESTU\_FECHANACIMIENTO -> EDAD

- PERIODO

- ESTU\_DEPTO\_RESIDE

- ESTU\_AREARESIDE

- ESTU\_TITULOOBTENIDOBACHILLER

- ESTU\_VALORMATRICULAUNIVERSIDAD

- ESTU\_PAGOMATRICULA

- ESTU\_COMOCAPACITOEXAMENSB11

- ESTU\_SEMESTRECURSA

- FAMI\_EDUCACIONPADRE

- FAMI\_EDUCACIONMADRE

- FAMI\_ESTRATOVIVIENDA

- FAMI\_TIENEINTERNET

- FAMI\_TIENESERVICIOTV

- FAMI\_TIENECOMPUTADOR

- FAMI\_TIENECONSOLAVIDEOJUEGOS

- ESTU\_HORASSEMANATRABAJA

- GRUPOREFERENCIA

- ESTU\_PRGM\_DEPARTAMENTO

- ESTU\_METODO\_PRGM

- INST\_CARACTER\_ACADEMICO

- ESTU\_INSE\_INDIVIDUAL

- ESTU\_NSE\_IES

Para cada variable se le realizo su respectiva imputación de valores nulos para los que tenían nulos. Esto fue a través de la Moda (para valores categóricos) y la media (para valores no categóricos).

Además, para cada variable se analizó su correlación con la variable objetivo y se generó su grafico de dispersión para analizar la distribución de los datos respecto a PUNT\_GLOBAL.

Se opto por dos estrategias para la transformación de datos, hacer dummies con los valores categóricos y realizar agrupaciones con los valores categóricos en nuevas categorías creadas. El modelo que mejor funciono fue el que se hizo con dummies y no con las agrupaciones.

Todo este proceso se puede ver en el repo en la carpeta tarea2 en el archivo indexSinAgrupaciones y indexConAgrupaciones

## Regresión

El modelo que mejor lo realizo saco los siguientes datos:

* R²: 0.33869430595889135
* MAPE: 18783856825415.688
* MSE: 369.2286036600235

Se observo que el modelo funciona ligeramente mejor con Test que con Train, esto debido a que la puntuación sacada en Kaggle siempre era un poco superior al MSE obtenido con los datos de training.

## KNN

Se realizo pruebas con 5, 10, 20, 30, 40, 50 y 60. En donde a partir de 30 el modelo no mejoro más. Y estos fueron los MSE obtenidos:

MSE para 5 vecinos: 443.95408178527623

MSE para 10 vecinos: 409.4955544587035

MSE para 20 vecinos: 397.5644615453648

MSE para 30 vecinos: 395.9491958412689

MSE para 40 vecinos: 396.6272564053666

MSE para 50 vecinos: 398.4814865416255

MSE para 60 vecinos: 399.9821943688282

El mejor número de vecinos es 30 con un MSE de 395.9491958412689

Lo observado es que el modelo de Regresión funciono mejor que la regresión y que consume mas entrenar el modelo KNN frente a Regresión.

## GBM

MSE en entrenamiento: 381.40471831154787

MAPE en entrenamiento: 8066752201627.534

MSE en prueba: 386.80937356163315

MAPE en prueba: 20583606131073.535