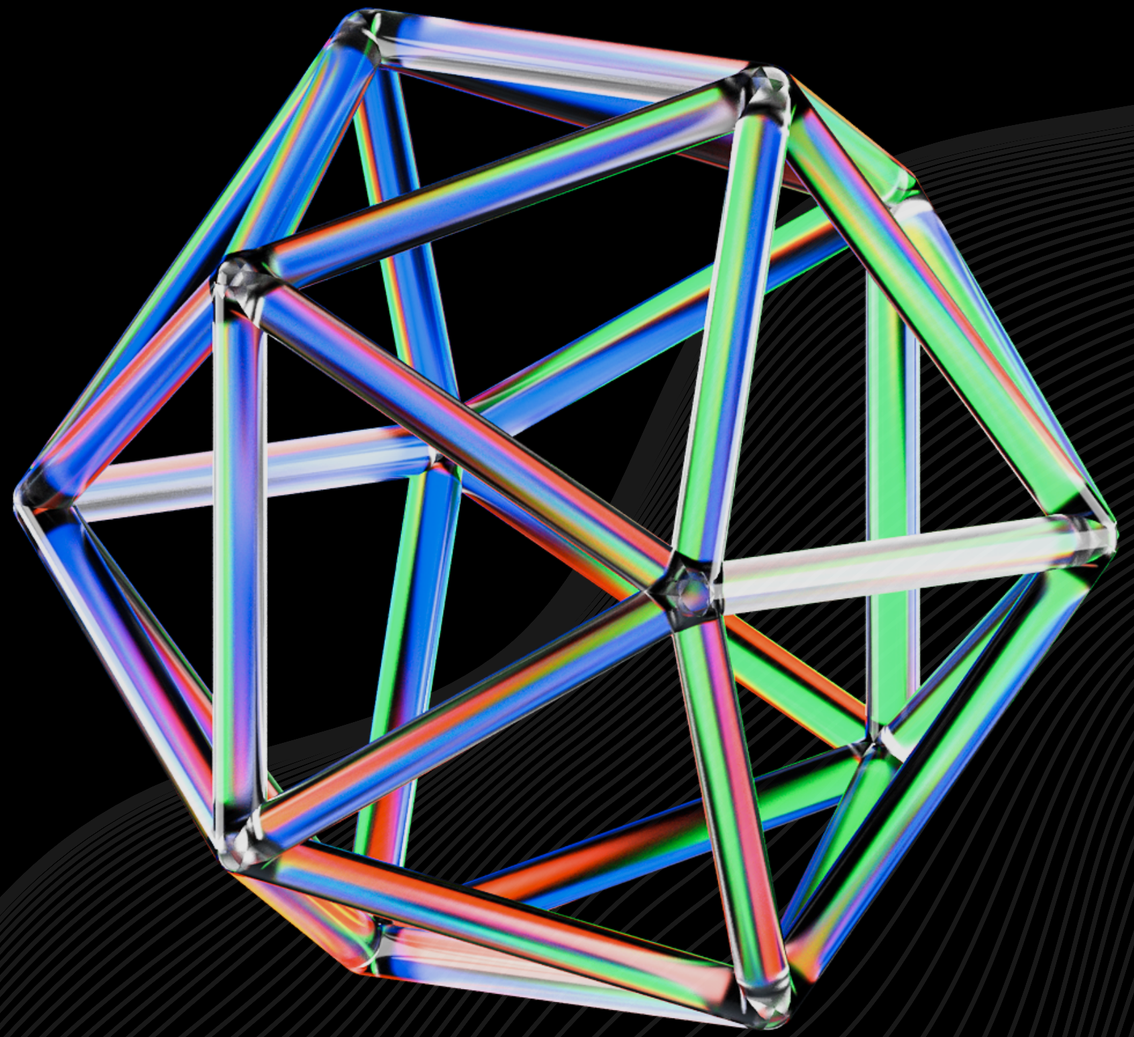
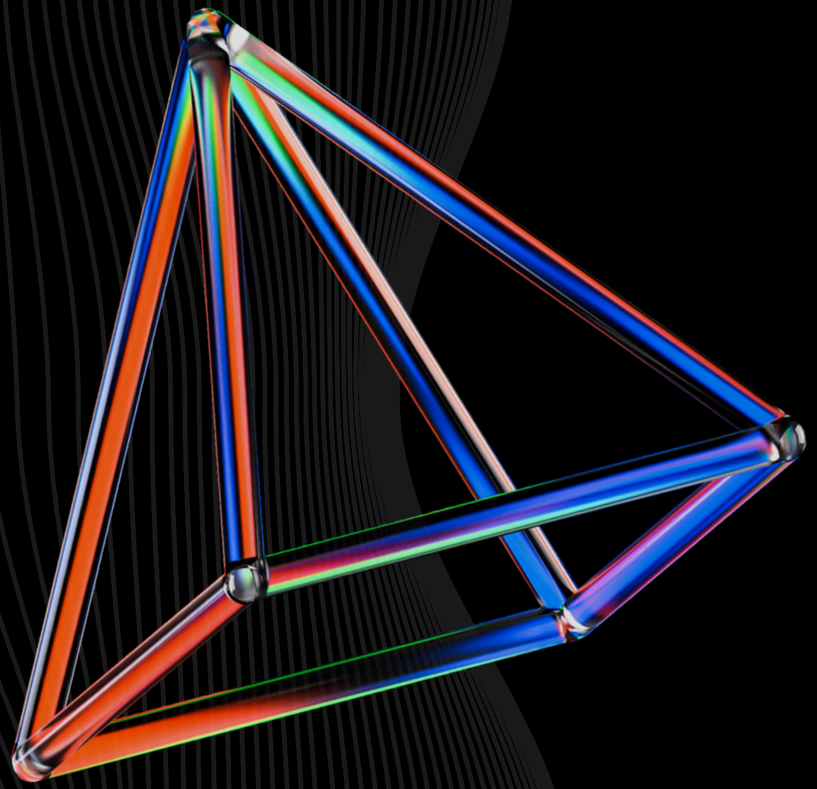
ANALITICA PREDICTIVA - ITBA

IMDB - AVG RATING PRED

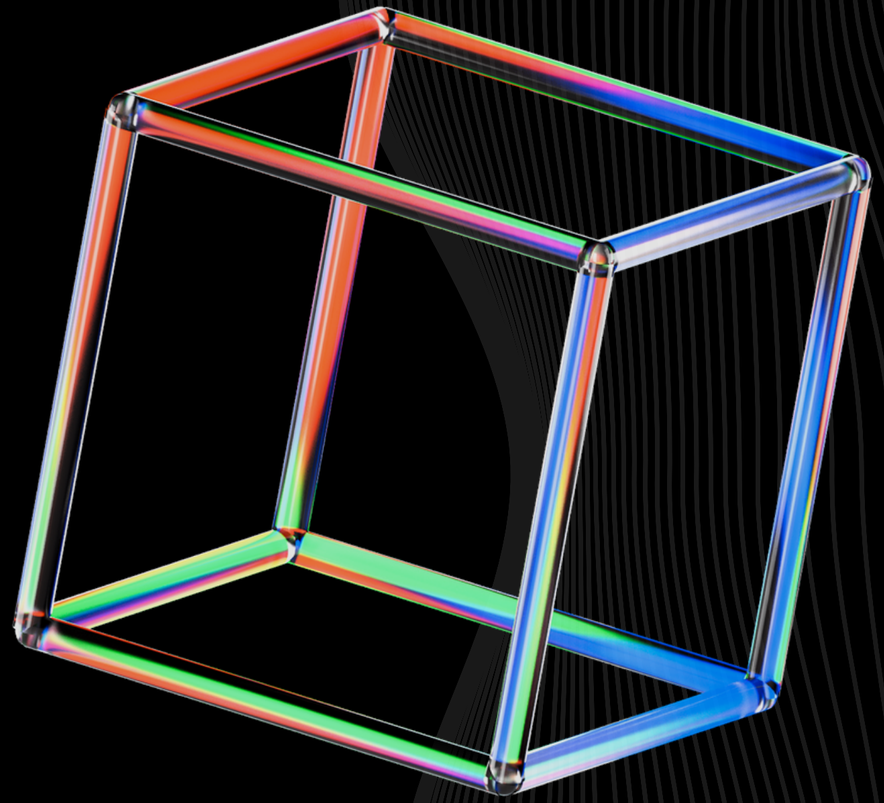


2022 - AGUSTIN ALFIE

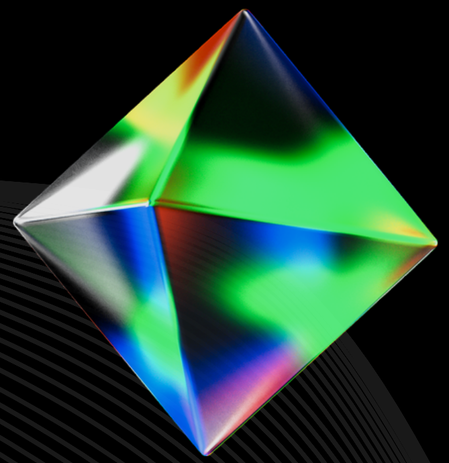


Índice

Introducción



Base de datos



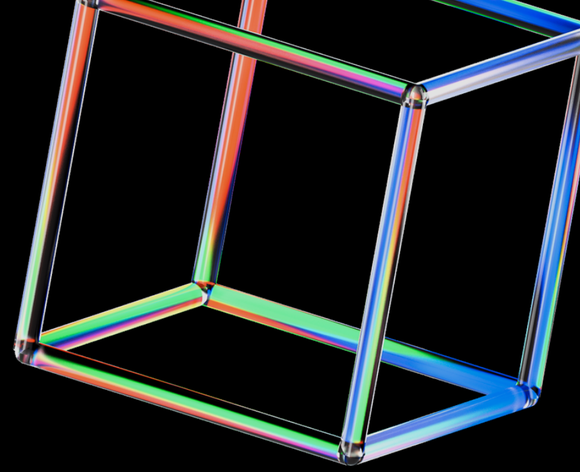
Registros

- 977541 registros
- Sin registros duplicados

Variables

- 31 variables
  - 8 categoricas
  - 10 numericas
  - 13 texto

Separacion teorica de la base



1

Variables con pocos o casi ningun registros vacio

2

Variables con maximo 65% de registros vacias

3

Varibles con 90% o mas de registros vacios

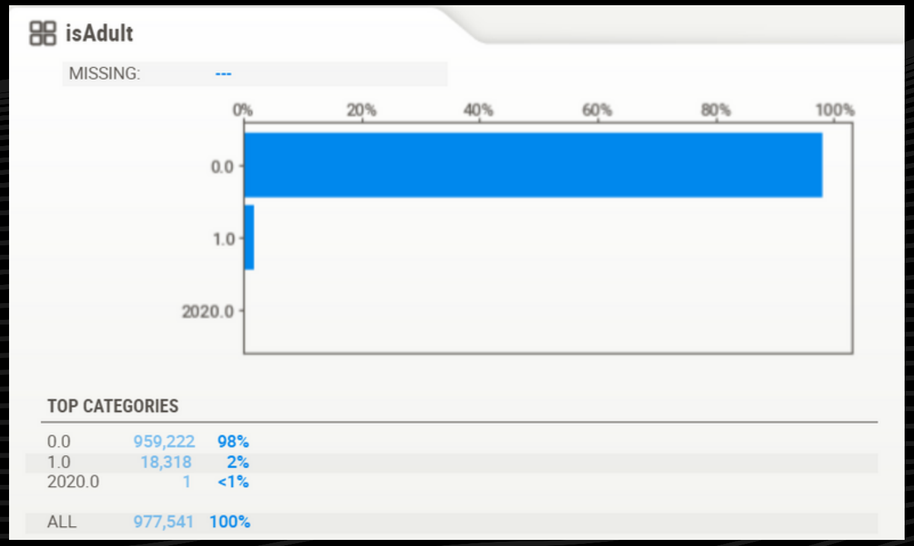
Análisis exploratorio

NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

runtime	0.15
runtimeMinutes	-0.12
ordering	-0.11
revenue	0.10
startYear	0.10
popularity	0.09
budget	0.05
episodeNumber	-0.04
seasonNumber	-0.03
numVotes	0.02
endYear	0.01
Id	-0.00

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

titleType	0.37
original_language	0.10
language	0.10
status	0.10
isAdult	0.06
adult	0.01
video	0.00
isOriginalTitle	0.00



Muchas variables con valores faltantes, por ejemplo language, 62% de valores faltantes y a su vez, la variable con mayor frecuencia es el 0, lo cual no trae mucha informacion al respecrto de idiomas,. ejemplos como estos hay muchos, en budget sucede lo mismo, 95% de valores nulos y el 77% de los datos aparece con un budget en cero. Esto nos obliga a asumir disntas cosas.

Asumciones

Langugae

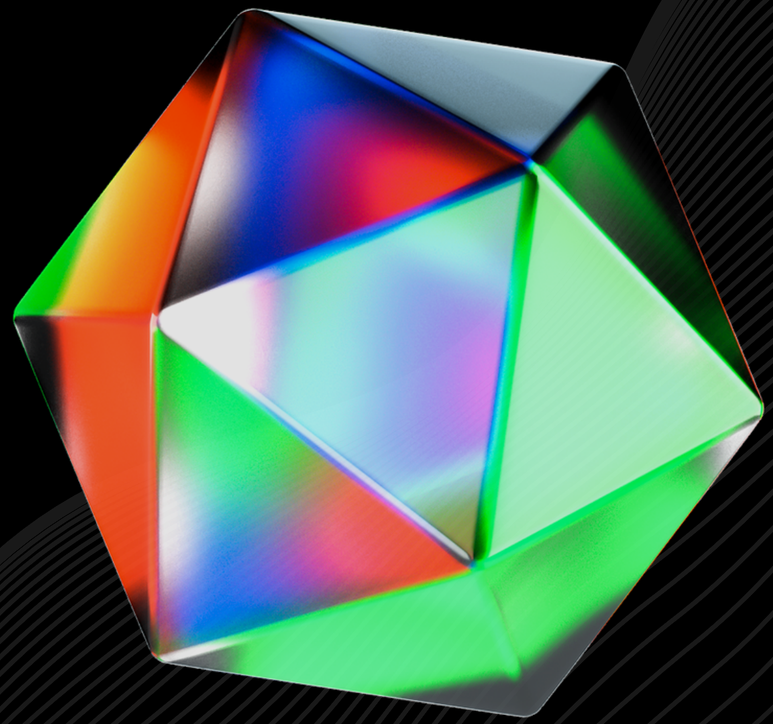
Los registros en la variable language se consideran como la moda, ingeles, se crea una varibale dumy de ingeles y se compara un ingles vs resto

Status

La variable status toma valores como released, production , post production, etc; los valores nulos se consideran como released, ya que se cree que la mayor parte de las peliculas/ series ya fueron estrenadas

Original title

Se considera a todas las peliculas con registros nulos en la variable original title como 1, titulos originales



Transformaciones Nulos y outliers

Generos

Escritores

Directores

idioma

es serie

Pais origen

title type

release_date

isadult

startyaer

runtime

revenue

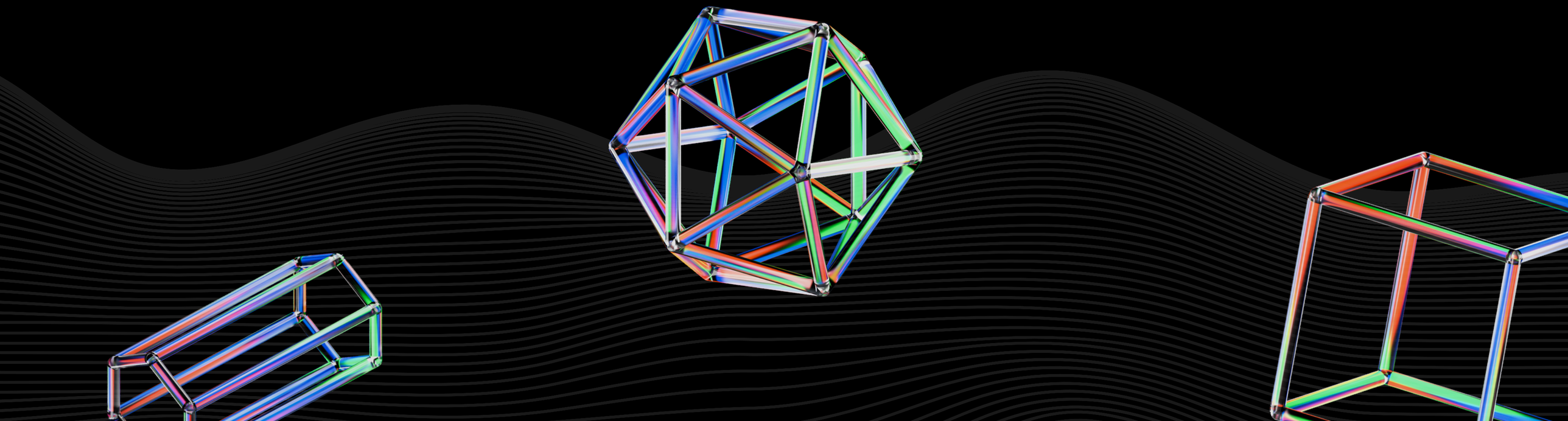
runtimeMinutes

ordening

Train test

Armado de dos train test split diferentes, uno para los modelos que no aceptan valores nulos y otros para los modelos que lo aceptan.

Variables eliminadas de la base sin NAs, revenue, runtime, anio, runtimeMinutes. popularity, epsiodeNumber



Modelos

Regresion lineal

r2 = 0.17
RMSE = 1.27
t = 1.1s

Random Forest

Random forest sin random search:
r2 = 0.29
RMSE= 1,17
t = 2m

Random search
r2 = 0.31
RMSE= 1.19
t= 14 m

Catboost

Catboost sin random search:
r2 = 0.34
RMSE= 1.3
t= 12s

Random search:
r2 = 0,37
RMSE = 1,12
t = 8 m

Xgboost

XgBoost sin random search:
r2 = 0.384
RMSE= 1,1
t = 1M

Random search
r2 = 0.386
RMSE= 1.10
t = 13 m

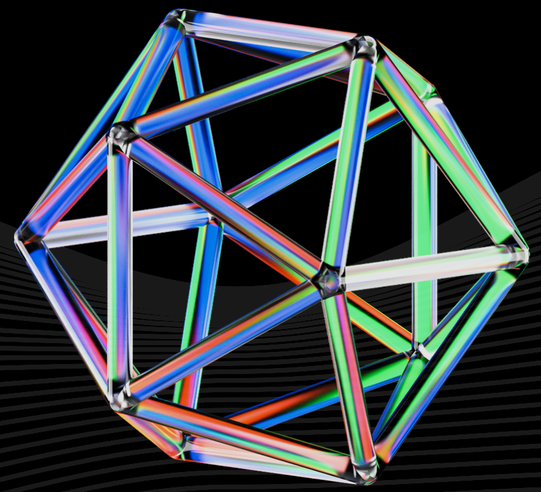
KNN

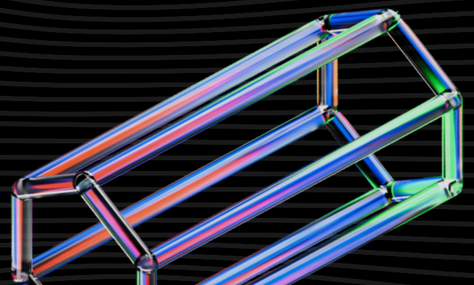
k = 30
r2 = 0.21
RMSE = 1.24
t = 17m

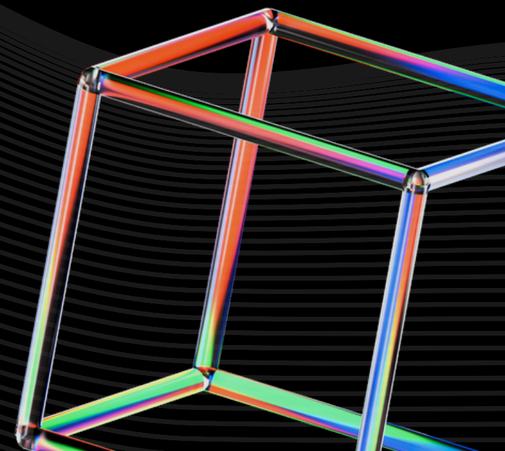
Modelo ganador

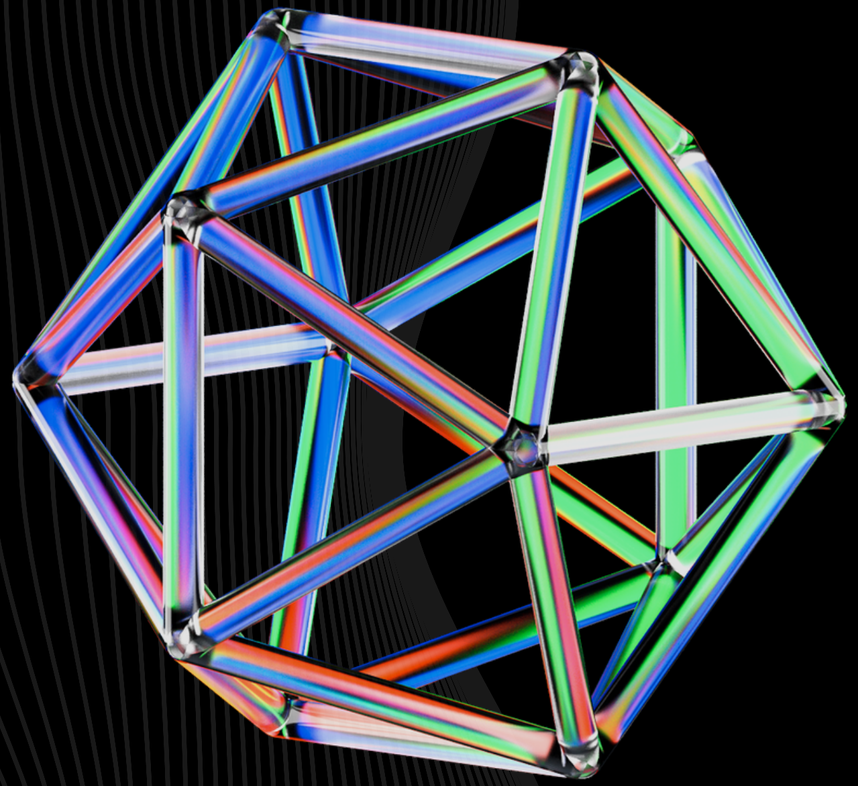
XgBoost:
'reg_alpha': 1,
'random_state': 0,
'n_estimators': 100,
'min_child_weight': 9,
'max_depth': 13,
'leraning_rate': 0.5,
'gamma': 0.4,
'colsample_bytree': 0.7

Random search
r2 = 0.386
RMSE= 1.10
t = 13 m









GRACIAS

2022 - Agustin Alfie