

# KAGGLE COMPETITION

## OBJETIVO

Esta competición trata de predecir el precio de un diamante en función de sus características.

Datasets:

- train.csv: contiene los precios para entrenar el modelo
- predict.csv: sin precios para realizar la predicción



## Steps

### I. Data Cleaning & Model Exploration

- Eliminar la columna de ID
- Cut, color and clarity: Convertir a valores numéricos (Label Encoder vs Diccionario)
- Explorar primeros modelos con sklearn y elegir los mejores en función del RMSE y R2

### II. Extra Trees Regressor

- ExtraTreesRegressor(n\_estimators=1800, min\_samples\_split=10, max\_depth= 50)
- RMSE: 538.47
- R2: 0.981557

### III. Hist Gradient Boosting Regressor

- HistGradientBoostingRegressor(loss='least\_squares', max\_depth=150, min\_samples\_leaf=2)
- RMSE: 544.39
- R2: 0.980244

### IV. Random Forest Regressor

- RandomForestRegressor(max\_depth=10, max\_features=10, min\_samples\_leaf=20, min\_samples\_split=15, n\_estimators=50)
- RMSE: 629.77
- R2: 0.973514

### V. H2O

- H2OAutoML(max\_runtime\_secs=1000, sort\_metric='RMSE')
- Model: StackedEnsemble (RMSE: 533.132)
- Model: GBM\_3 (RMSE: 539.697)

## Conclusiones

El modelo ganador es el generado de manera automática por H2O:

- StackedEnsemble
  - RMSE: 533.132