# KAGGLE COMPETITION

13/11/2023

82.05 - Análisis Predictivo - Secundo Parcial

Fanny LATRON (65998)

- **28 variables and 977 541 average rates**

- **Missing values:**

| | |
|---|---|
| averageRating | 0 |
| numVotes | 0 |
| titleType | 0 |
| isAdult | 0 |
| startYear | 0 |
| endYear | 0 |
| runtimeMinutes | 0 |
| genres_x | 2 |
| directors | 0 |
| writers | 0 |
| seasonNumber | 539298 |
| episodeNumber | 539298 |
| ordering | 606918 |

| | |
|---|---|
| language | 606918 |
| attributes | 606918 |
| isOriginalTitle | 606918 |
| adult | 930171 |
| budget | 930171 |
| genres_y | 930171 |
| original_language | 930183 |
| popularity | 930172 |
| production_companies | 930172 |
| production_countries | 930172 |
| revenue | 930172 |
| runtime | 930383 |
| status | 930242 |
| tagline | 953696 |
| video | 930172 |

- **seasonNumber and episodeNumber** → if part of 'movie', 'video', 'tvSpecial', 'tvMovie', 'videoGame', then 0 season and 1 episode

  → if part of 'tvSeries', 'tvEpisode', 'short', 'tvMiniSeries', 'tvShort', then 1 season and 7 episodes

- **ordering** → filled with the mean

- **genres_x** → dropped the 2 lines with missing values

- **directors and writers** → took the 1st name

- **startYear and endYear** → if endYear = 0, then endYear= startYear

- **First Model : decision Tree**

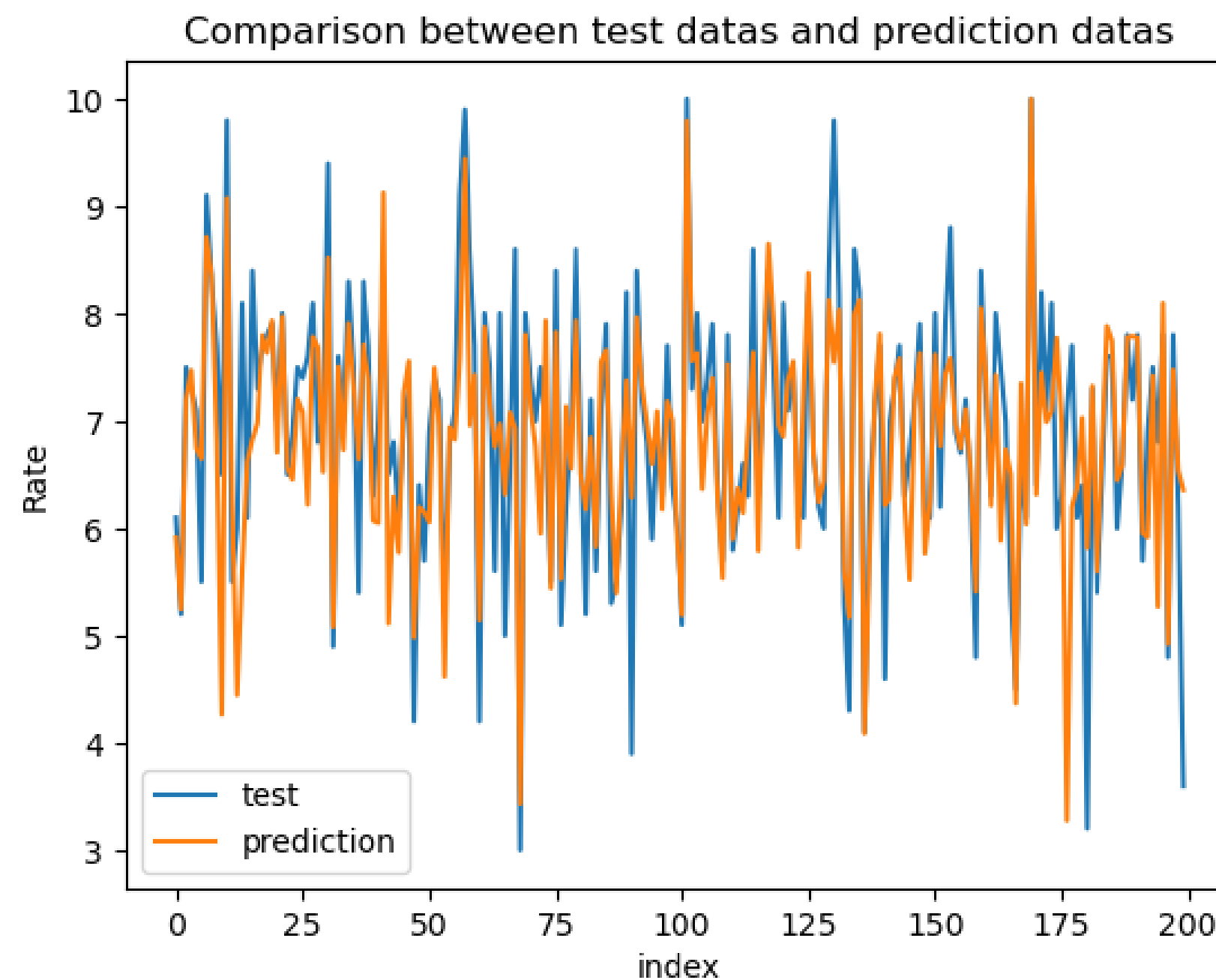  → Parameters: max_depth =5, nb_estimators = 30, learning_rate = 0.07

  → $R^2 \approx 0.2$

- **Other Model : SVM**

● **Random Forest**

→ Parameters: max_depth =35, nb_estimators = 150

→ $R^2$ = 0.49657



Comparison between test datas and prediction datas

# LIMITS

- execution time

- parameters not optimal

- $R^2$ still far from 1

# Thanks !