ITBA
Instituto Tecnológico de Buenos Aires

KAGGLE COMPETITION

13/11/2023

82.05 - Análisis Predictivo - Secundo Parcial

Fanny LATRON (65998)

CLEANING

- 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

CLEANING

- **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

MODELS

- **First Model : decision Tree**

  → Parameters

  → $R^2 \approx 0.2$

- **Other Model : SVM**

FINAL MODEL

- **Random Forest**

CONCLUSION

- mainly categorical variables
- no significant correlation between the variables, so no redundant variables
- some variables have more weight in the explanation of the variable to be explained : GenHlth, HighBP, DiffWalk, HighChol, Age
- possibility of adding a column (for example, individuals with a family member suffering from diabetes)

Thanks !