Methodology

* The 3 classifiers used- **Decision tree, logistic regression**
* Ensemble pipeline
* Other models considered
* Hyper-parameter tuning- **criterion="entropy", max\_depth=12**

Overview

* Objective- **Identify the most relevant demographic attributes and develop a model to predict the items for a user. Check if the new model improved the prediction or not**.
* Methodology-

Dataset

* How many features- **33 features**
* Size of the dataset- **100 000**
* Multiple files- **3**
* What kind of data – numerical or character**- numerical, categorical, DateTime**
* Balanced or imbalanced – what is the distribution- **imbalanced**
* Distribution of Training set, validation set, testing set- **60-20-20 percentage**
* Missing data and Preprocessing challenges- **Release date has missing data, we removed null values and converted it to years. We have done one-hot encoding for categorical variables. We had dependent variables as one-hot-encoded, so we have flattened our output variable to single column**

Feature Engineering Techniques

* Features removed**- "movie id","timestamp", "IMDb URL","video release date", "user id", “movie title”**
* Feature creation- **converted release date to years**
* Feature ranking- **Not needed**
* Class imbalance treatment- **Not needed**
* Any other

Results

* Table for the evaluation metric for each ML technique used
* Plot of the curves
* Conclusion