Raport z projektu

06-DUMAUI0 2022/23/SZ

# Cel projektu

Celem projektu było stworzenie modelu, który przewiduje ocenę filmu (0.0-10.0) na podstawie danych o filmie (rok, obsada, gatunki, liczba głosów, kategoria wiekowa).

# Dane

Dane pochodzą z oficjalnego datasetu IMDb (linki: <https://www.imdb.com/interfaces/> i <https://datasets.imdbws.com/> ).

Wykorzystane zostały pliki title\_basic\_data.tsv, title\_principals\_data.tsv oraz title\_raing\_data.tsv.

Plik title\_rating\_data.tsv zawiera dane z ocenami filmów, odrzucone zostały z niego filmy zawierające mniej niż 100 ocen.

Plik title\_basic\_data.tsv zawiera informacje of filmach, wykorzystane kolumny to: idfilmu, kategoria wiekowa filmu, rok, gatunki. Kolumna gatunki została podzielona na 3 osobne kolumny, a następnie metodą hot ones encoding na dane numeryczne.

Plik title\_principals\_data.tsv zawiera informacje o obsadzie filmowej. Na początku korzystałem z niego, ale później został odrzucony, żeby poprawić model.

Po odrzuceniu obserwacji odstających, uzyskano 319 565 przykładów, które podzielono na zbiór uczący (255 652 przykłady) i zbiór testowy (63 913 przykładów).

# Modele

W projekcie porównano działanie 2 modeli:

* Regresja liniowa
* Sieci neuronowe wykorzystujące regresję liniową

# Ewaluacja

Do ewaluacji wykorzystano metrykę mean\_squared\_error. Wyniki ewaluacji przedstawia poniższa tabelka:

|  |  |
| --- | --- |
| Model | Mean\_squared\_error |
| Regresja linowa | 1.2341117124384207 |
| Sieć neuronowa | 1.5849324067433674 |

# Wnioski

Oba modele nie dostarczyły wystarczająco dobrych wyników, spowodowane jest to najpewniej źle przygotowanym datasetem, niestety próbowałem go modyfikować na różne sposoby ale żaden nie dostarczył zadowalających wyników. Normalna regresja liniowa dostarczyła lepszy wyników niż sieć neuronowa.