

# Евалуација на алгоритми за препорака на филмови

Мила Ѓуровска 231116

## 1. Вовед

Системите за препораки се критична компонента на модерните апликации, од стриминг сервиси до е-комерц платформи. Целта на оваа домашна е да се евалуираат различни алгоритми за филтрирање во контекст на препоруки на филмови, користејќи го датасетот MovieLens.

Анализираните алгоритми вклучуваат:

- Spark ALS (Alternating Least Squares)
- Surprise SVD (Singular Value Decomposition)
- Surprise KNN User-Based (K-Nearest Neighbors - базирано на корисници)
- Surprise KNN Item-Based (K-Nearest Neighbors - базирано на предмети)

## 2. Методологија

### 2.1 Датасет

Користен е датасетот MovieLens, кој содржи оценки на филмови од различни корисници. Датасетот обезбедува реалистична основа за обучување и тестирање на препорачни системи.

### 2.2 Алгоритми за евалуација

**Spark ALS:** Дистрибуиран алгоритам за матрична факторизација, оптимизиран за обработка на големи датасети преку паралелна обработка.

**Surprise SVD:** Класичен алгоритам за матрична факторизација базиран на сингуларна декомпозиција, широко користен во системите за препорачување.

**KNN User-Based:** Алгоритам кој препорачува филмови врз основа на подобност со други корисници со слични оценки.

**KNN Item-Based:** Алгоритам кој препорачува филмови врз основа на сличност на филмовите, без разлика на сличностите меѓу корисници.

### 2.3 Метрики за евалуација

За евалуација на перформансите на алгоритмите, користени се следните метрики:

- **RMSE (Root Mean Square Error)**: Ја мери просечната разлика помеѓу предвидените и вистинските оценки. Помала вредност значи подобра прецизност.
- **MAE (Mean Absolute Error)**: Ја мери просечната апсолутна разлика. Исто како RMSE, помала вредност е подобра.

## 3. Резултати

Следната табела ги прикажува добиените резултати од експериментите:

Алгоритам	RMSE	MAE
Spark ALS	0.920019	0.728176
Surprise SVD	0.938046	0.739379
Surprise KNN User-Based	1.022054	0.809111
Surprise KNN Item-Based	1.036686	0.822684

Табела 1: Перформанси на алгоритми за препорачување на филмови

**Spark ALS**: Постигна најниска RMSE вредност од 0.920019 и најниска MAE од 0.728176, демонстрирајќи најдобра точност во предвидување на оценки. Дистрибуираната природа на алгоритмот овозможува ефикасна обработка на големи датасети.

**Surprise SVD**: Вториот најдобар резултат со RMSE од 0.938046 и MAE од 0.739379. Овој алгоритам е компактен и релативно брз.

**Surprise KNN User-Based**: Покажа полоши резултати во споредба со претходните алгоритми, со RMSE од 1.022054 и MAE од 0.809111. Ова е бидејќи методот на К-најблизу соседи може да бидат чувствителни на квалитетот на сличностните мерила.

**Surprise KNN Item-Based**: Ги покажа најслабите резултати со RMSE од 1.036686 и MAE од 0.822684. Item-based пристапот може да биде неоптимален кога филмовите немаат јасна категоризација или сличност.

## 4. Заклучок

Оваа евалуација демонстрира дека Spark ALS е најдобриот избор за препорачање на филмови во овој контекст. Неговата комбинација на висока точност и скалабилност го прави идеален за производни примени. Иако KNN методите имаат своја улога и се широко користени, резултатите ја покажуваат супериорноста на матричната факторизација за овој тип на проблем.