

# Movies Recommendations Based on Collaborative Filtering.

ROLL NO	NAME
BCA-30	Kshiteeja Jadhav
BCA-69	Jyoti
BCA-70	Ajinkya Gaikwad

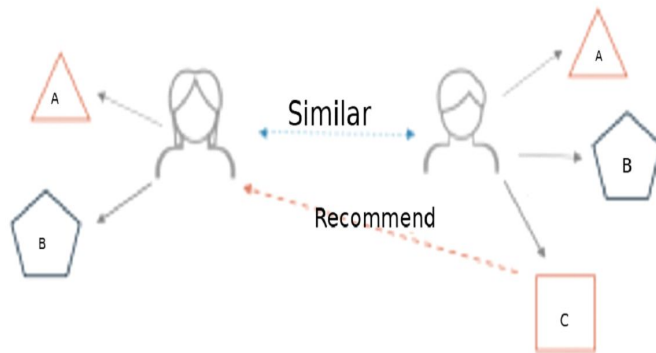
# What is a recommender system?

A recommender system is a simple algorithm whose aim is to provide the most relevant information to a user by discovering patterns in a dataset. The algorithm rates the items and shows the user the items that they would rate highly. An example of recommendation in action is when you visit Amazon and you notice that some items are being recommended to you or when Netflix recommends certain movies to you. They are also used by Music streaming applications such as Spotify and Deezer to recommend music that you might like.

Below is a very simple illustration of how recommender systems work in the context of an e-commerce site.

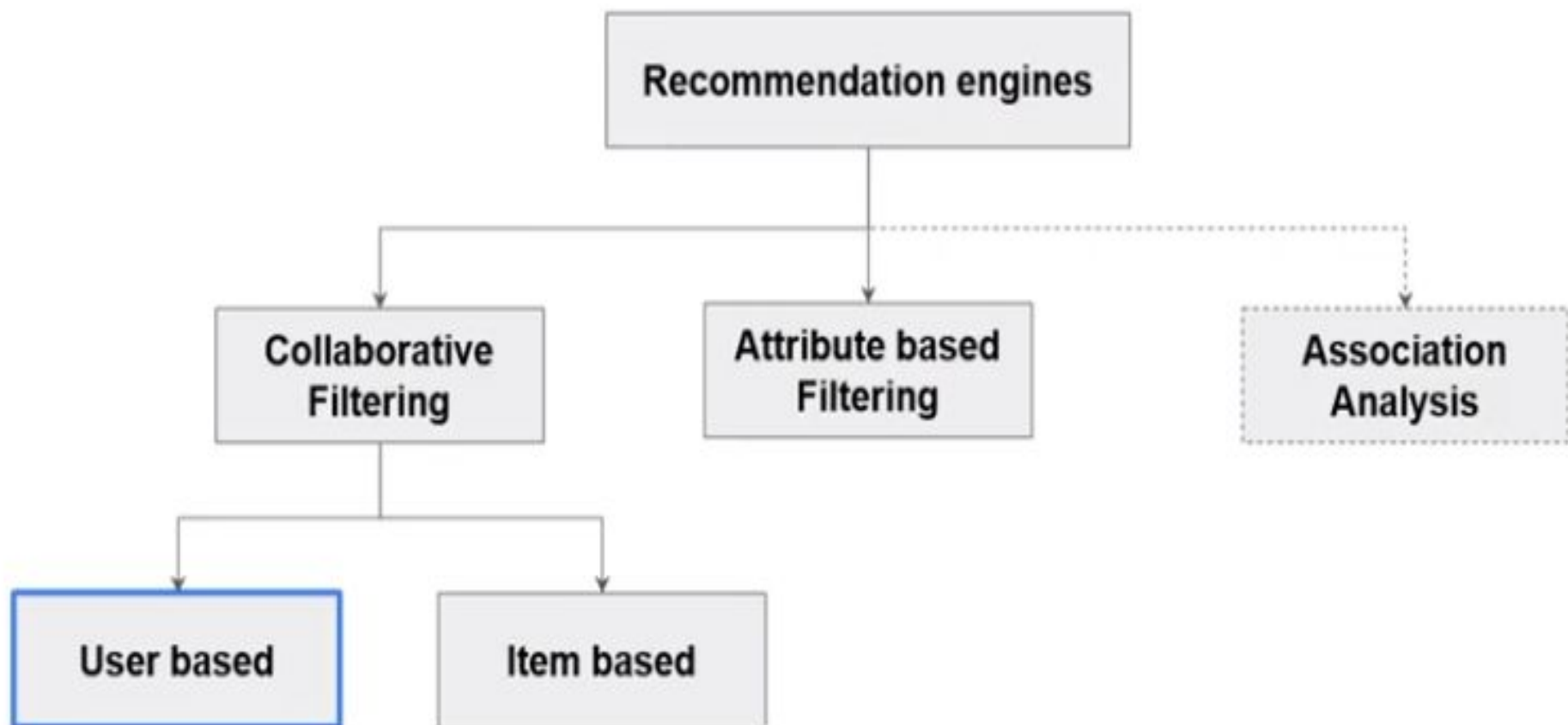
# What is a recommender system?

Two users buy the same items A and B from an e-commerce store. When this happens the similarity index of these two users is computed. Depending on the score the system can recommend item C to the other user because it detects that those two users are similar in terms of the items they purchase.



# Different types of recommendation engines.

The most common types of recommendation systems are content-based and collaborative filtering recommender systems. In collaborative filtering, the behavior of a group of users is used to make recommendations to other users. The recommendation is based on the preference of other users. A simple example would be recommending a movie to a user based on the fact that their friend liked the movie. There are two types of collaborative models Memory-based methods and Model-based methods. The advantage of memory-based techniques is that they are simple to implement and the resulting recommendations are often easy to explain. They are divided into two:



# Different types of recommendation engines:

**User-based collaborative filtering:** In this model, products are recommended to a user based on the fact that the products have been liked by users similar to the user. For example, if Derrick and Dennis like the same movies and a new movie come out that Derrick like, then we can recommend that movie to Dennis because Derrick and Dennis seem to like the same movies.

- **Item-based collaborative filtering:** These systems identify similar items based on users' previous ratings. For example, if users A, B, and C gave a 5-star rating to books X and Y then when a user D buys book Y they also get a recommendation to purchase book X because the system identifies book X and Y as similar based on the ratings of users A, B, and C.

# Different types of recommendation engines:

Content-based systems use metadata such as genre, producer, actor, musician to recommend items say movies or music. Such a recommendation would be for instance recommending Infinity War that featured Vin Diesel because someone watched and liked The Fate of the Furious. Similarly, you can get music recommendations from certain artists because you liked their music. Content-based systems are based on the idea that if you liked a certain item you are most likely to like something that is similar to it.

# Datasets Used:

We are going to use the Movie Lens Data Set. This dataset was put together by the Group lens research group at the University of Minnesota. It contains 1, 10, and 20 million ratings. Movie lens also has a website where you can sign up, contribute reviews and get movie recommendations.

The Datasets are as follows:

- movies.csv
  - movieid
  - title
  - genre
- ratings.csv
  - userId
  - movieid
  - Ratings
  - timestamp



Ratings for a  
few items



Ratings for all the  
items



Utility matrix: Users x Item  
ratings

kNN Algorithm to find  
similar users

# K-Nearest Neighbours Algorithm:

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method proposed by Thomas Cover used for classification and regression.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If  $k = 1$ , then the object is simply assigned to the class of that single nearest neighbor.

# K-Nearest Neighbours Algorithm:

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, normalizing the training data can improve its accuracy dramatically.

## Dataset

### **MovieLens**

100,000  
ratings from  
1,000 users  
and 1,700  
movies

## Parameters

k in kNN:  
**80**

## Tool



**rapidminer**

# Movies.csv:

Windows taskbar and system tray are visible on the left. The main application window is titled "RapidMiner Studio Educational 9.7.002 @ DESKTOP-U9VJLR2". The menu bar includes File, Edit, Process, View, Connections, Settings, Extensions, and Help. The toolbar contains icons for file operations and a "Views" dropdown set to "Results".

The "Results" view is active, showing a table of movie data. The table has columns: Row No., movielid, title, and genres. The data is filtered to show 9,124 examples.

Row No.	movielid	title	genres
1	1	Toy Story (19...	Adventure Ani...
2	2	Jumanji (1995)	Adventure Ch...
3	3	Grumpier Old...	Comedy Ro...
4	4	Waiting to Ex...	Comedy Dra...
5	5	Father of the ...	Comedy
6	6	Heat (1995)	Action Crime ...
7	7	Sabrina (1995)	Comedy Ro...
8	8	Tom and Huc...	Adventure Ch...
9	9	Sudden Deat...	Action
10	10	GoldenEye (1...	Action Advent...
11	11	American Pre...	Comedy Dra...
12	12	Dracula: Dea...	Comedy Horror
13	13	Balto (1995)	Adventure Ani...
14	14	Nixon (1995)	Drama
15	15	Cutthroat Isla...	Action Advent...
16	16	Casino (1995)	Crime Drama

ExampleSet (9,124 examples, 0 special attributes, 3 regular attributes)

The "Repository" panel on the right shows the project structure. The "Data" folder under "Movie\_Recommendation\_2 (Local)" is expanded, showing "movies" and "ratings" files.

# Ratings.csv:

Windows taskbar: Desktop, 11:02 PM Monday 10/26/2020, 4 notifications

RapidMiner Studio Educational 9.7.002 @ DESKTOP-U9VJLR2

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

Result History ExampleSet (//Movie\_Recommendation\_2/Data/ratings)

Open in Turbo Prep Auto Model

Filter (100,004 / 100,004 examples): all

Row No.	userid	rating	movielid	timestamp
1	1	2.500	31	1260759144
2	1	3	1029	1260759179
3	1	3	1061	1260759182
4	1	2	1129	1260759185
5	1	4	1172	1260759205
6	1	2	1263	1260759151
7	1	2	1287	1260759187
8	1	2	1293	1260759148
9	1	3.500	1339	1260759125
10	1	2	1343	1260759131
11	1	2.500	1371	1260759135
12	1	1	1405	1260759203
13	1	4	1953	1260759191
14	1	4	2105	1260759139
15	1	3	2150	1260759194
16	1	2	2193	1260759198

ExampleSet (100,004 examples, 2 special attributes, 2 regular attributes)

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Local Repository (Local)
  - Connections
  - data
  - processes
  - pivot calculation (10/26/20 10:54 PM)
- movie\_recommend\_1 (Local)
- Movie\_Recommendation\_2 (Local)
  - Connections
  - Data
    - movies (10/15/20 6:27 PM - 1.5 M)
    - ratings (10/15/20 6:28 PM - 3.5 M)
  - Process
    - dmw\_project (10/26/20 10:56 PM - )
- Temporary Repository (Local)
  - Connections
  - Cross Distances Temp 47274 (10/15/20 6:27 PM - 1.5 M)
  - Select Attributes Temp 42469 (10/15/20 6:27 PM - 1.5 M)
  - Split Data Temp 49301 (10/15/20 6:27 PM - 1.5 M)
  - Split Data Temp 62270 (10/15/20 6:27 PM - 1.5 M)
  - Subprocess Temp 3606 (10/26/20 10:56 PM - )

# Process:

Windows taskbar: Desktop, 11:03 PM Monday 10/26/2020, 4 notifications

RapidMiner Studio Educational 9.7.002 @ DESKTOP-U9VJLR2

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

**Repository**

- Import Data
- Movie\_Recommendation\_2
  - Connections
  - Data
    - movies (10/15/20 6:27 PM)
    - ratings (10/15/20 6:26 PM)
  - Process
    - dmw\_project (10/26/20 10:5)
- Temporary Repository (Local)
  - Connections
  - Cross Distances Temp 47
  - Select Attributes Temp 424

**Operators**

Search for Operators

- Blending (82)
- Attributes (48)
- Examples (11)
- Filter (2)
- Sampling (6)
- Sample

[Get more operators from the Marketplace](#)

**Process**

Process

```
graph LR; inp --> Movies; Movies --> Join; Join --> Rating2[Rating (2)]; Rating2 --> SelectAttributes; SelectAttributes --> AllRolesSet; AllRolesSet --> SplitData; SplitData --> MovieID1[MovieID]; SplitData --> MovieID2[MovieID (2)]; MovieID1 --> UserKNN[User k-NN]; MovieID2 --> ApplyModel; ApplyModel --> res;
```

**Parameters**

Process

logverbosity init

logfile

[Show advanced parameters](#)

[Change compatibility \(9.7.002\)](#)

**Recommended Operators**

**Problems**

No problems found

Message	Fixes	Location
---------	-------	----------

Activate Windows  
Go to Settings to activate Windows.

Recommendations Update



# Process Output:

Windows taskbar: Desktop, 11:03 PM, Monday, 10/26/2020

RapidMiner Studio Educational 9.7.002 @ DESKTOP-U9VJLR2

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

Result History ExampleSet (MovieID (2))

Open in Turbo Prep Auto Model

Filter (10,000 / 10,000 examples): all

Row No.	rating	title	genres	userId	movieId	prediction
1	3	Toy Story (19...	Adventure Ani...	67	1	3.770
2	3.500	Toy Story (19...	Adventure Ani...	134	1	4.139
3	3	Toy Story (19...	Adventure Ani...	149	1	3.827
4	2	Toy Story (19...	Adventure Ani...	176	1	3.043
5	5	Toy Story (19...	Adventure Ani...	219	1	4.080
6	2	Toy Story (19...	Adventure Ani...	220	1	3.843
7	5	Toy Story (19...	Adventure Ani...	242	1	4.308
8	5	Toy Story (19...	Adventure Ani...	268	1	3.938
9	4	Toy Story (19...	Adventure Ani...	272	1	3.233
10	3	Toy Story (19...	Adventure Ani...	283	1	3.715
11	4.500	Toy Story (19...	Adventure Ani...	345	1	3.892
12	5	Toy Story (19...	Adventure Ani...	383	1	3.262
13	3.500	Toy Story (19...	Adventure Ani...	456	1	3.431
14	1.500	Toy Story (19...	Adventure Ani...	457	1	3.083
15	5	Toy Story (19...	Adventure Ani...	459	1	3.812
16	4	Toy Story (19...	Adventure Ani...	502	1	4.117

ExampleSet (10,000 examples, 5 special attributes, 1 regular attribute)

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Local Repository (Local)
  - Connections
  - data
  - processes
  - pivot calculation (10/26/20 10:54 PM)
- movie\_recommend\_1 (Local)
- Movie\_Recommendation\_2 (Local)
  - Connections
  - Data
    - movies (10/15/20 6:27 PM - 1.5 M)
    - ratings (10/15/20 6:28 PM - 3.5 M)
  - Process
    - dmw\_project (10/26/20 10:56 PM)
- Temporary Repository (Local)
  - Connections
  - Cross Distances Temp 47274 (10/26/20 10:56 PM)
  - Select Attributes Temp 42469 (10/26/20 10:56 PM)
  - Split Data Temp 49301 (10/15/20 5:00 PM)
  - Split Data Temp 62270 (10/15/20 5:00 PM)
  - Subprocess Temp 3606 (10/26/20 10:56 PM)



# Performance:

Windows taskbar: Desktop, 11:03 PM Monday 10/26/2020, 4 notifications.

RapidMiner Studio Educational 9.7.002 @ DESKTOP-U9VJLR2

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

**Repository**

- Import Data
- Movie\_Recommendation\_2 (Local)
  - Connections
  - Data
    - movies (10/15/20 6:27 PM)
    - ratings (10/15/20 6:26 PM)
  - Process
    - dmw\_project (10/26/20 10:5)
  - Temporary Repository (Local)
    - Connections
    - Cross Distances Temp 47
    - Select Attributes Temp 424

**Operators**

Search for Operators

- Item Rating Prediction
- Recommender Perform
  - Model Application (4)
  - Performance Evalua
    - Performance (Iter
    - Performance (Ra

**Process**

Process

Process Diagram:

- Input (inp) connects to **Movies** (file icon).
- Movies** connects to **Join** (join icon).
- Join** connects to **Rating (2)** (file icon).
- Rating (2)** connects to **Select Attributes** (table icon).
- Select Attributes** connects to **All Roles Set** (table icon).
- All Roles Set** connects to **Split Data** (funnel icon).
- Split Data** connects to **MovieID** (table icon).
- Split Data** connects to **MovieID (2)** (table icon).
- MovieID** connects to **User k-NN** (lightbulb icon).
- MovieID (2)** connects to **Apply Model** (lightbulb icon).
- User k-NN** connects to **Performance** (percentage icon).
- Apply Model** connects to **Performance**.
- Performance** has outputs: res, res, per, eva.

**Parameters**

**Performance (Performance (Rating Prediction))**

Min Rating: 1

Range: 4

**Recommended Operators**

**Problems**

No problems found

**Message**

**Fixes**

**Location**

Activate Windows. Go to Settings to activate Windows. [Show advanced parameters](#)

# Performance Output:

The screenshot displays the RapidMiner Studio interface. The main window shows the 'PerformanceVector (Performance)' tab, which displays the RMSE (Root Mean Square Error) value of 0.916. The left sidebar contains a 'Result History' panel with a 'Performance' section. The right sidebar shows a 'Repository' panel with a tree view of the project structure.

**PerformanceVector (Performance)**

Criterion

- RMSE
- MAE
- NMAE

**RMSE**

RMSE: 0.916

**Repository**

- Training Resources (connected)
- Samples
- Community Samples (connected)
- Local Repository (Local)
  - Connections
  - data
  - processes
  - pivot calculation (10/26/20 10:54 PM)
- movie\_recommend\_1 (Local)
- Movie\_Recommendation\_2 (Local)
  - Connections
  - Data
    - movies (10/15/20 6:27 PM - 1.5 M)
    - ratings (10/15/20 6:28 PM - 3.5 M)
  - Process
  - dmw\_project (10/26/20 10:56 PM - 10:56 PM)
- Temporary Repository (Local)
  - Connections
  - Cross Distances Temp 47274 (10/15/20 5:00 PM)
  - Select Attributes Temp 42469 (10/15/20 5:00 PM)
  - Split Data Temp 49301 (10/15/20 5:00 PM)
  - Split Data Temp 62270 (10/15/20 5:00 PM)
  - Subprocess Temp 3606 (10/26/20 10:56 PM)

# Performance Output:

The screenshot displays the RapidMiner Studio interface. The main window shows the 'Performance' tab for a model named 'PerformanceVector (Performance)'. The 'Criterion' list on the left includes RMSE, MAE (selected), and NMAE. The 'MAE' value is displayed as 0.702. The 'Repository' panel on the right shows a tree structure of data sources, including 'Local Repository (Local)' and 'Temporary Repository (Local)'. The 'Local Repository' contains 'Connections', 'data', 'processes', and 'pivot calculation'. The 'Temporary Repository' contains 'Connections', 'Cross Distances Temp 47274', 'Select Attributes Temp 42469', 'Split Data Temp 49301', 'Split Data Temp 62270', and 'Subprocess Temp 3606'. The 'Connections' folder in the 'Temporary Repository' is expanded, showing these sub-processes. The 'Subprocess Temp 3606' is highlighted. The 'Connections' folder in the 'Local Repository' is also expanded, showing 'data' and 'processes'. The 'data' folder is expanded, showing 'pivot calculation'. The 'processes' folder is expanded, showing 'movie\_recommend\_1' and 'Movie\_Recommendation\_2'. The 'Movie\_Recommendation\_2' is expanded, showing 'Connections', 'Data', 'Process', and 'Temporary Repository'. The 'Data' folder is expanded, showing 'movies' and 'ratings'. The 'ratings' folder is expanded, showing 'ratings'.

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

Result History PerformanceVector (Performance)

Criterion

- RMSE
- MAE
- NMAE

Performance

Description

Annotations

MAE

MAE: 0.702

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Local Repository (Local)
  - Connections
  - data
  - processes
  - pivot calculation ( 10/26/20 10:54 PM)
- movie\_recommend\_1 (Local)
- Movie\_Recommendation\_2 (Local)
  - Connections
  - Data
    - movies ( 10/15/20 6:27 PM - 1.5 M)
    - ratings ( 10/15/20 6:26 PM - 3.5 M)
  - Process
    - dmw\_project ( 10/26/20 11:04 PM - 1.5 M)
  - Temporary Repository (Local)
    - Connections
    - Cross Distances Temp 47274 ( 10/15/20 6:27 PM - 1.5 M)
    - Select Attributes Temp 42469 ( 10/15/20 6:27 PM - 1.5 M)
    - Split Data Temp 49301 ( 10/15/20 6:27 PM - 1.5 M)
    - Split Data Temp 62270 ( 10/15/20 6:27 PM - 1.5 M)
    - Subprocess Temp 3606 ( 10/26/20 11:04 PM - 1.5 M)

# Performance Output:

The screenshot displays the RapidMiner Studio interface. The main window shows the 'PerformanceVector (Performance)' tab, which displays the 'NMAE' criterion with a value of 0.175. The left sidebar contains icons for 'Performance', 'Description', and 'Annotations'. The right sidebar shows the 'Repository' panel, which lists various data sources and processes. The 'Repository' panel is expanded, showing a tree structure of data and processes. The 'Data' folder is expanded, showing 'movies' and 'ratings' datasets. The 'ratings' dataset is selected, showing its details. The 'Processes' folder is also expanded, showing a list of processes including 'dmw\_project', 'Cross Distances Temp 47274', 'Select Attributes Temp 42469', 'Split Data Temp 49301', 'Split Data Temp 62270', and 'Subprocess Temp 3606'. The 'dmw\_project' process is selected, showing its details. The 'Performance' tab in the main window shows the 'NMAE' criterion with a value of 0.175. The 'Description' and 'Annotations' tabs are also visible in the sidebar.

Performance

Criterion

RMSE

MAE

NMAE

NMAE: 0.175

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Local Repository (Local)
  - Connections
  - data
  - processes
    - pivot calculation (10/26/20 10:54 PM)
- movie\_recommend\_1 (Local)
- Movie\_Recommendation\_2 (Local)
  - Connections
  - Data
    - movies (10/15/20 6:27 PM - 1.5 M)
    - ratings (10/15/20 6:28 PM - 3.5 M)
  - Process
    - dmw\_project (10/26/20 11:04 PM - 1.5 M)
- Temporary Repository (Local)
  - Connections
  - Cross Distances Temp 47274 (10/15/20 6:27 PM - 1.5 M)
  - Select Attributes Temp 42469 (10/15/20 6:28 PM - 3.5 M)
  - Split Data Temp 49301 (10/15/20 6:28 PM - 3.5 M)
  - Split Data Temp 62270 (10/15/20 6:28 PM - 3.5 M)
  - Subprocess Temp 3606 (10/26/20 11:04 PM - 1.5 M)

Thank You

