Journal of Computer Science

Original Research Paper

High Accurate Multicriteria Cluster-Based Collaborative Filtering Recommender System

Zhila Yaseen Taha and Sadegh Abdollah Aminifar

Department of Computer Science, Soran University, Soran, Erbil, Kurdistan Region, Iraq

Article history
Received: 08-07-2022
Revised: 26-10-2022
Accepted: 24-11-2022

Corresponding Author:
Sadegh Abdollah Aminifar
Department of Computer Science, Soran University,
Soran, Erbil, Kurdistan Region,
Iraq
Email: sadegh.aminifar@soran.edu.iq

Abstract: A Recommender System (RS) is one of the appropriate answers by researchers for alleviating the problem of customers' overload with information from an internet source. In RS, users' evaluation of the items assists in quickly determining the user's preferences and removing the choosing overhead for the user the next time they search. However, using user ratings for an item based on multiple aspects has given researchers little attention to recommending an object accurately. This research proposes a model-based collaborative filtering movie recommendation system based on user ratings across various criteria. The proposed systems' model was established using clustering and classification methods, such as k-means, k-modes, and Multinomial logistic regression. The proposed work consists of two different steps; first, clustering mode; after clustering the dataset with k-means, k-modes were used to re-cluster it into many sub-clusters for search domain reduction purposes. The second is classification mode; the Multinomial Logistic Regression (MLR) model was created to predict the closest cluster class to newly active users. Since the MLR is one of the probabilistic models, it is more accurate than cluster methods in the prediction process. The proposed approach uses distance indicators, such as modified Mahalanobis distance and Euclidean distance, to measure the similarity between new active and in-group users. The evaluation of the proposed methodology was done using the MAE and Silhouette scores. Different values of Silhouette score were achieved in this study for a different number of features and clusters, with the best k-means score being 0.822. Based on multicriteria Yahoo!! Movie Dataset, the MAE result indicates that this study is better than the existing one.

Keywords: Collaborative Filtering, Clustering-Based CF, k-Means, k-Modes, Multinomial-Logistic Regressions

Introduction

A recommender system is a technique used to filter collected information over internet users to perform two primary tasks: First, anticipating user ratings and then suggesting a new item for them that they have not examined before (Xue *et al*., 2019). In recent years, the RS has attracted the attention of analysts and they have established many recommendation models with the same goal of benefiting both the customer and the service provider organization, as well as dealing with the problem of data overabundance. One of the most frequent techniques to construct an RS is the collaborative filtering model, which offers recommendations to customers based on the choice of other users (Marzuki *et al*., 2014). The most crucial step in CF is to make a reliable forecast for a target user by combining concepts that are comparable to their own. In most circumstances, CF follows a two-step process: First, it looks for customers with similar ratings and utilizes their ratings to produce a suggestion for that new customer (Wasid and Ali, 2018).

To locate similar users, some studies, for instance (Yassine *et al*., 2021; Ahuja *et al*., 2019; Aminifar and Marzuki, 2013a; 2013b), employed model-based CF, such as k-means clustering; some of them used hybrid methods like k-nearest neighbors and k-means clustering. Our system consists of two different steps; first, clustering mode; after clustering the Dataset, to perform search domain reduction, the dataset has been re-clustered into more than one sub-cluster using k-modes. Then, to achieve higher clustering accuracy and avoid complex problems, the powerful feature extraction technique has been applied called Principal Component Analysis (PCA), which reduced dataset size and memory usage and visualized the Dataset more efficiently. The sub-clustering process achieved dual goals:

Science Publications

- Reducing the search area and hence the search time
- It further reduces the dissimilarity across users within the same cluster and groups users who are highly similar

As the second step of the proposed strategy, a new classification model called MLR has been created. The MLR forecasts which cluster is most likely to contain new active users. The multinomial logistic regression uses a SoftMax function to predict the nearest group, which results in high accuracy in prediction. Finally, different distance indicators, such as modified Mahalanobis distance and Euclidean distance measure used to predict the closest supercluster and discover the top N nearest users to the new active- users who have rated at least one movie. The main contributions of this studya are:

1. The implementation of k-modes clustering as a sub-clustering strategy to reduce dissimilarity between users in the same group
2. Innovatively classify the unsupervised dataset using the result of k-means clustering
3. Using different methods for measuring the similarity between users such as Mahalanobis and Euclidean similarity indicators

Related Work and Materials

This section is divided into two sub-sections. The first sub-section is a comprehensive review of studies on the two forms of CFRSs (memory-based and model-based CFRSs). In the second sub-section, effective methods used in our proposed approach have been discussed.

In general, Recommender Systems proved to be the most reliable due to their ability to filter relevant data in a reasonable time possible. However, many setbacks face RS methods. To address these setbacks and offer a thriving RS with the primary criteria, the researchers used a range of unique methodologies in their investigations; the sparsity and Scalability issues of the CFRS method have been addressed in a research article (Mustaqeem *et al*., 2020). To this end, after performing the data preprocessing steps over the row data, they managed missing values, irrational data, and repetitive records by applying the Numerical cleaner filtering technique. They then addressed the search space domain problem and improved user similarity between users by benefiting from the sub-clustering approach.

The research author (Wasid and Ali, 2018) handled the multidimensionality using a sub-clustering approach. Likewise, (Sun and Dong, 2017) used enhanced k-means clustering and a time impact factor matrix to track the level of user interest drift in the class and, more precisely, estimate an item's rating. Their method employs item clustering, which predicts user interest levels indirectly using user ratings. In work (Yassine *et al*., 2021), they employed user demographics like gender and age to classify users into different profiles. They employed k-means clustering to cluster movies based on the film genre; after separating the most viewed movie with users of the same gender, they applied SVD to generate an efficient recommendation.

Furthermore, research (Ahuja *et al*., 2019; Abd and Aminifar, 2022a) uses many different ML methods and techniques, such as k-nearest neighbor and k-means clustering. They used the WCSS method with the elbow approach to calculate and discover the correct number of clusters. The research author (Thakkar *et al*., 2019) devised a strategy to decrease the prediction error. To this end, they merged predictions from two CF methods, named user-based CF (UbCF) and Item-based CF (ICF), over Multiple Linear Regression (MLR) and Support Vector Regression (SVR) methods.

Also, the Cold start and data sparsity are two CF concerns that received attention in work (Natarajan *et al*., 2020). They developed RS using the LOD model as a new similarity indicator metric that integrates the enhanced PCC and PICSS, which find items similar to the target item. The information about the new entities gathered utilizing the DBpedia cloud and the cold start issue is solved. They solved the data sparsity issue using the enhanced matrix factorization with LOD. Another work proposed by Ajaegbu (2021) to improve ICF-RS The authors, in their work, also focused on the cold-start issue by enhancing the similarity metric used in the traditional ICFRS method. Accurately depicting user preference has been the focus of work (Hu *et al*., 2020). For this purpose, they established the matrix factorizations method with multiplex implicit user feedback, which forecasts unknown user ratings and solves data sparsity problems. Moreover, to assess customer emotion from the text of user reviews and doctor feature distribution, (Zhang *et al*., 2017) introduced a novel healthcare RS technique. to this end, they employed sentiment analysis and topic modeling in combination with the hybrid matrix factorization model. Additionally, the author (Qian *et al*., 2019), research, suggested an enhanced emotion-aware RS using UBCF and IBCF techniques. They outlined behavior-change theories, theoretical components, and health promotion as three groups related to healthcare systems. The test result showed that the suggested strategy significantly raises prediction ratings and increases suggestion accuracy. The RS has been used to recommend movies to the user in research (Chen, 2021). The authors created the Profile-Based (PB) module to accumulate the attributes of each film. Each individual developed positive and negative profiles using the most extensive and minor ratings; based on these profiles. They obtained the actual ratings through a CF module and the expected ratings through the UbCF module. Also. Furthermore, research (Aminifar, 2020; Liu *et al*., 2017) enhanced the user similarity computation metric; they altered the coverage method by adding the logarithm. They used precision and recall to evaluate their enhanced UBCF.

Materials

Collaborative Filtering

Collaborative filtering is a prominent machine learning strategy for making predictions and recommendations based on system users with similar tastes. The purpose of CFRS is to assist each new user in retrieving a reliable prediction by mixing concepts that are similar to their own.in most cases, the CF performs two steps: First is discovering users that have an equal rating to the new active user and then utilizing their ratings to produce a suggestion for that new user. To calculate user similarity, the CF uses a cosine, Pearson correlation coefficient, Euclidean distance, and a variety of other indications.

K-Means Clustering

It is a partitioning approach in which datasets are iteratively separated into k number of clusters a (Kant *et al*., 2018; Jader *et al*., 2022). The standard k-means clustering method for multicriteria rating datasets is as follows:

- Determine the ideal number of k for the Dataset
- Pick k numbers from the Dataset at random to use as a centroid
- K-means uses the Euclidean distance indicator. First, calculate the distance between the centroids and users. Then, assign each user to the center that is nearest to them
- Calculate the mean of each cluster and perform cluster reassignment iteratively until reaching the convergence criteria
- Terminate the algorithm after convergence criteria

K-Modes Clustering

The k-modes approach is a modified version of the k-means method used for clustering large categorical datasets-the modification of k-modes is in three aspects (Kuo *et al*., 2021):

- K-modes count the dissimilarity between the centroid and each data point instead of the distance
- K-modes use mode instead of means to find a new centroid
- K-modes find modes by using the frequency-based method

K-modes apply a similar k-means theory but measure the dissimilarity between data points and centroid rather than the distance between them. K-modes also use mode to indicate the new cluster Centre rather than means. Removing the restriction of numerical data and the ability to scale to massive data sets are the main advantage of using k-modes (Hamzah *et al*., 2017; Jader and Aminifar, 2022a; 2022b).

Multinomial Logistic Regression

It is a modified alternative to the logistic regression model that directly supports the prediction and defines multi-class labels. To be more specific, to estimate the likelihood that an input example corresponds to each known class label. Multinomial logistic regression: A logistic regression model customized to learn and predict probability distribution. Multinomial logistic regression uses a cross-entropy function to compute the distance between the SoftMax function's generated probability and the one-hot-encoding matrix (Brownlee, 2021; Aminifar, 2006).

Mahalanobis Distance Indicator

This method computes the distance between the cluster centroid and each point of data using the variance of the Dataset, as shown in Eq. (1):

$$MD\left(M_{ui}, kc_i\right) = \sqrt{\frac{\left(kc_i - Mui,\right)^2}{\left(\sigma i, j\right)^2}} \tag{1}$$

whereas, M_{ui} is a multicriteria user rating, kc_i is the centroid, and MD (M_{Ui}, kc_i) is a Mahalanobis distance between each centroid and new-active user ratings. σi, j is a variance of each feature of all of each user in the same cluster.

Euclidean Distance Indicator

The Euclidean distance is the best indicator for estimating the distance points along a straight line. Equation (2) shows the formula for calculating the similarity between n-dimensional points (Marjai *et al*., 2021):

$$D = \sqrt{\sum_{i=1}^{n}\left(u_{i,n} - c_{i,n}\right)^2} \tag{2}$$

where:
$u_{i,n}$ = A user rating for item *n*
$c_{i,n}$ = A ratings of cluster centroid *c* for item *n*

Proposed Approach

The entire work is programmed in Python using k-means, k-modes, and Multinomial Logistic Regression. The system's implementation is divided into several sub-sections, as illustrated in Fig. 1. these include the following:

- Dataset collection
- Preparing and preprocessing the dataset
- Feature reduction step using PCA
- Supermodel using k-means
- Sub-Clustering K-mode approach
- Classification and recommendation
- Evaluation

Dataset

In this study, the Yahoo movie dataset has been utilized, which consists of two files (movies and ratings), with 1k multicriteria ratings provided by webscop@verizonmedia.com. The summary and some statistical information about the Dataset are illustrated in Table 1.

Preparing and Preprocessing the Dataset

After the data collection process, the data preparation and preprocessing are significant steps to enhance the quality of the available Dataset. To this end, the obtained dataset has been prepared to meet the study strategy requirement by removing all additional and pointless columns from the two files, such as timestamps, tags, etc., and replacing each movie id with its title. Moreover, a python module has been used to merge all pertinent features from the two CSV files into a single Data Frame. Furthermore, we looked at additional statistical information to better compresence the available dataset, for instance, the minimum and maximum ratings, the number of users, the number of rows and null rows, the amount of memory used, etc. Once the Data Frame has been created, the data preprocessing step was performed using the transformation and dimensional reduction process. For this purpose, the Min-Max scaler function to normalize and transform the rating features has been applied, as illustrated in Table 2.

Super-Sub Clustering Model

Super-Clustering Approach

Clustering is one of the model-based CFRSs that groups unsupervised data based on how similar the points in the dataset are to one another. The clustering technique has been utilized to identify comparable users in the unsupervised dataset based on their rating scores for each movie. the clustering process was done in two steps: Called supercluster and sub-cluster.

K-means clustering was employed in the supercluster phase, to group comparable users and detect user interest from the available dataset based on user ratings given to each movie. This was done by performing the following steps:

Step 1: Importing the required libraries and classes from the Scikit Learn package

Step 2: Determining the value of k as an initial step in the k-means algorithm and evaluating the centroids of datasets using the Elbow approach. After fitting our Dataset and testing it with centroids ranging from 2 to 20, the optimal value for k was 6, with inertia values (sum of distance among each point in a dataset and its assigned cluster) of nearly 2701.1, as shown in Fig. 2

Step 3: Construct the k-means model. The completed model randomly selects the centroids based on the number of k

Step 4: Use the Fit_Predict () method to fit the Dataset with constructed model after choosing the centroids. The k-means algorithm uses a distance indicator metric to compute the distance between centroids and each user. Let's use a multidimensional (movie rating based on multi aspects) dataset as an instance and assume that, $M_{ui} = \{x_1, x_2, x_3, x_4\}$ and $M_{Ci} = \{c_1, c_2, c_3, c_4\}$, the Eq. (3) shows calculation distance between M_{ui} and M_{CI}:

$$\frac{D\left(M_{Ui}, M_{Ci}\right)}{\sqrt{\left(c_1 - x_1\right)^2 + \left(c_2 - x_2\right)^2 + \left(c_3 - x_3\right)^2 + \left(c_4 - x_4\right)^2}} \tag{3}$$

where, M_{ui} is a multi-aspect rating of user I and M_{CI} is a multi-aspect rating of cluster-centroid I.

Step 5: Iteratively process the centroid computation and user assignment to the nearest centroid until the convergence condition. Accordingly, after more than 16 iterations, it reached convergence criteria. The result of clustering the dataset is shown in Table 4

Step 6: Plotting the dataset's clustering results, as seen in Fig. 3

Sub-Clustering Approach

The aim of decreasing the search space is to simplify and reduce the processing complexity, which reduces the time it takes to recommend items. However, the dataset's clustering technique did not adequately reduce the search domain. Each cluster class still has a sizeable number of users and movies. Therefore, it was necessary to re-cluster each obtained class. Thus, the k-modes were the best choice to use as a sub-clustering strategy. The goal of adopting the k-mode was not only to reduce the search area but also to reduce the dissimilarity among users in the same cluster. To this end, we took the following steps to complete the sub-clustering phase:

- Importing the required python libraries
- Separating the data of each supercluster class and generating a new data Frame for each of them
- Computing the cost and found the best k for each cluster class separately using the elbow, as shown in Table 3
- Visualizing the outcome of the values of k and its cost, as shown in Fig. 4
- Constructing the k-modes model. The constructed model randomly selects the centroids first based on the number of k for each sub-cluster class
- Using the Fit_Predict () method to fit the dataset with constructed k-modes model after choosing the centroids and re-assigning users to the nearest cluster. To obtain convergence criteria, the number of iterations was fixed manually to 6 for each sub-cluster class

The result of the sub-clustering is shown in Tables 4 to 9.

Classification and Recommendation

Now, it is necessary to determine which cluster the newly active user who has already rated at least one movie will be placed in.?

Who is the top N nearby users to new-active users based on the rated movie.? and which movie should be recommended to that new user? The best answer for this problem was to create a new model using multinominal logistic regression as a classification algorithm. As previously stated, the dataset (unsupervised) was clustered into six groups based on their closeness during the supercluster phase: Now, each user rating has a label, either (0,1,2,3,4 or 5), As shown in Fig. 5. So, the dataset was treated as a supervised dataset. Thus, each rating is a feature and each cluster label is a label. Based on this assumption, a new model using the MLR method was created; for that, the supercluster data was divided into two parts (80 and 20%) for training and testing, respectively. The total number of train and test rows was (80668, 20168) rows.

Table 1: Summary of used dataset

Number of rows	100867
The number of users rated	610
Number of movies	9742
Rating-range	0 to 5
The average number of rated movies for each user	20
Number of null row-columns	0
All user Id count	100836.000000
Memory usage	4.6+ MB

Table 2: Result of applying PCA over the dataset

	PCA1	PCA2
1	0.165584	-0.053777
2	-0.093793 -	-0.193015
3	-0.170804	-0.382430
4	-0.071190	-0.027764
5	-0.535155	0.498268
6	0.286112	0.205141
7	-0.059023	0.126841
8	0.006279	0.105666
9	-0.509872	-0.023782
10	-0.154759	0.121149

Table 3: The result of calculating the cost of sub-clustering each supercluster

Number of sub-clusters	Cost of sup-clustering cluster 1	Cost of sup-clustering group 2	Cost of sup-clustering group 3	Cost of sup-clustering group 4	Cost of sup-clustering group 5	Cost of sup-clustering group 6
1	32428.0	67707.0	53879.0	37309.0	47905.0	36372.0
2	29496.0	58749.0	45593.0	35009.0	44315.0	33784.0
3	27515.0	53024.0	43806.0	30791.0	39862.0	29211.0
4	25658.0	49742.0	38597.0	29817.0	38430.0	27647.0
5	25460.0	50075.0	37880.0	28096.0	35081.0	26653.0
6	23880.0	45111.0	35637.0	27706.0	35290.0	25898.0
7	23010.0	43423.0	35439.0	26089.0	33808.0	24888.0
8	23669.0	43598.0	32785.0	25533.0	33336.0	24555.0
9	21953.0	42506.0	33022.0	25556.0	30122.0	24499.0

Table 4: Distributing the data from the first supercluster among each sub-cluster

Cluster 1	Number of rows	Number of unique users	Number of unique movies
First	6689	583	2781
Second	4459	531	2268

Table 5: Distributing the data from the second supercluster among each sub-cluster

Cluster 2	Number of rows	Number of unique users	Number of unique movies
First	6689	583	2781
Second	3057	522	1591
Third	4060	510	2238
Fourth	4359	530	2308
Fifth	4459	531	2781

Table 6: Distributing the data from the third supercluster among each sub-cluster

Cluster 3	Number of rows	Number of unique users	Number of unique movies
First	12673	605	3779
Second	7869	579	3046

Table 7: Distributing the data from the fourth supercluster among each sub-cluster

Cluster	Number of rows	Number of unique users	Number of unique movies
First	4721	497	2636
Second	4116	526	2306
Their	2519	433	1672
Fourth	1515	372	1152

Table 8: Distributing the data from the fifth supercluster among each sub-cluster

Cluster	Number of rows	Number of unique users	Number of unique movies
First	11014	574	4112
Second	4712	499	2524
Their	2466	413	1693

Table 9: Distributing the data from the sixth supercluster among each sub-cluster

Cluster	Number of rows	Number of unique users	Number of unique movies
First	7002	585	2535
Second	3009	517	1687

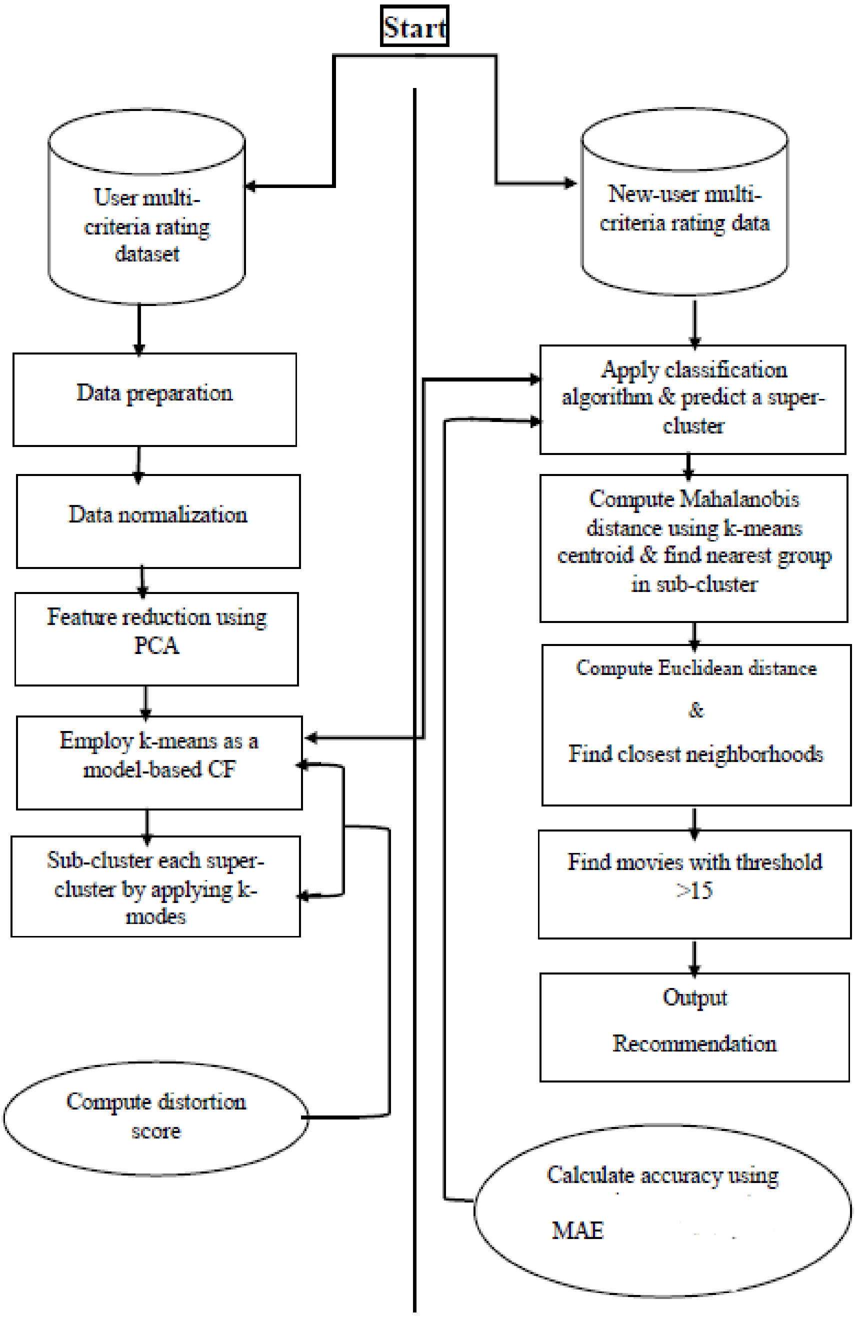

Start
User multi-criteria rating dataset
New-user multi-criteria rating data
Data preparation
Data normalization
Feature reduction using PCA
Employ k-means as a model-based CF
Sub-cluster each super-cluster by applying k-modes
Compute distortion score
Apply classification algorithm & predict a super-cluster
Compute Mahalanobis distance using k-means centroid & find nearest group in sub-cluster
Compute Euclidean distance
&
Find closest neighborhoods
Find movies with threshold >15
Output
Recommendation
Calculate accuracy using
MAE

Fig. 1: Cluster-based movie recommender system block diagram

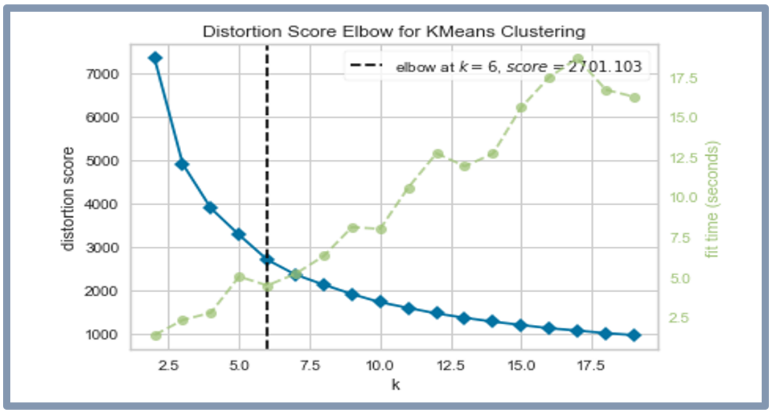


Fig. 2: Discovering the best possible value for k in k-means by implementing the elbow approach

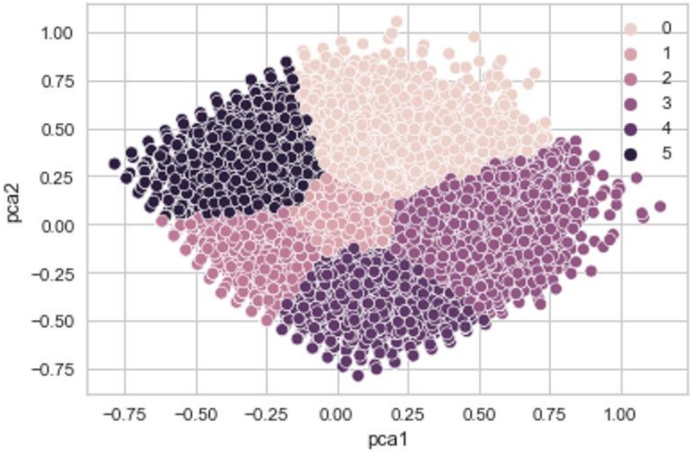


Fig. 3: The dataset distribution after applying k-means clustering

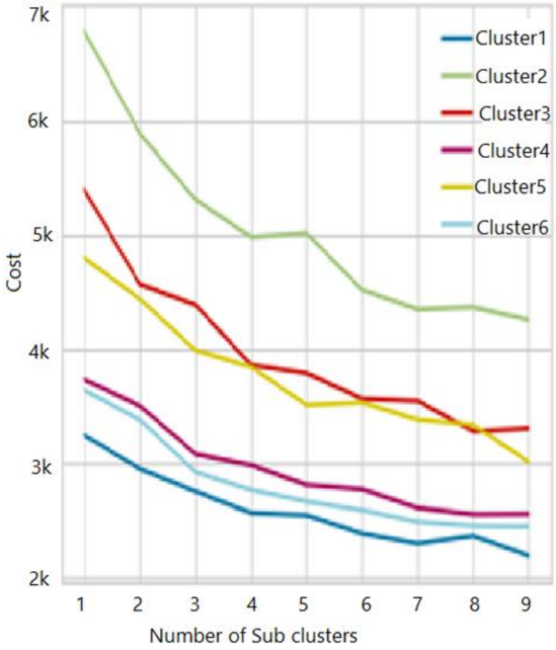
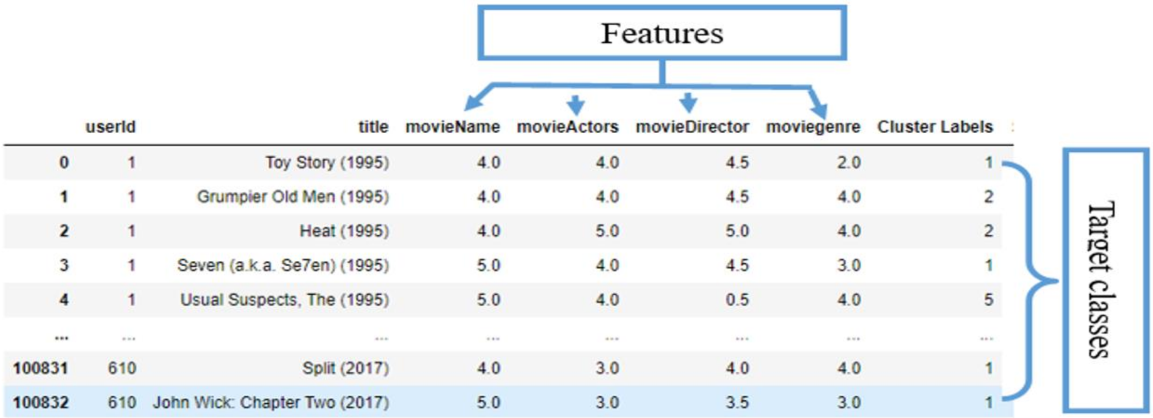


Fig. 4: The optimum k value for sub-clustering the supercluster



	userId	title	movieName	movieActors	movieDirector	moviegenre	Cluster Labels
0	1	Toy Story (1995)	4.0	4.0	4.5	2.0	1
1	1	Grumpier Old Men (1995)	4.0	4.0	4.5	4.0	2
2	1	Heat (1995)	4.0	5.0	5.0	4.0	2
3	1	Seven (a.k.a. Se7en) (1995)	5.0	4.0	4.5	3.0	1
4	1	Usual Suspects, The (1995)	5.0	4.0	0.5	4.0	5
...	...	...	...	...	...	...	...
100831	610	Split (2017)	4.0	3.0	4.0	4.0	1
100832	610	John Wick: Chapter Two (2017)	5.0	3.0	3.5	3.0	1

Fig. 5: The outcome of applying the k-means clustering algorithm to the dataset

To predict a class, MLR will first compute the probability for each label in the training phase using the SoftMax function, then make a prediction using the the-cross-entropy process. The first step is creating a linear-prediction model as illustrated in Eq. (4):

$$ML = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \tag{4}$$

Whereas, *ML*= multi-dimensional linear predictor model, *W* is the weight of each feature. B bias. The w (weight) and b (bias) are random numbers at thea first step, in our dataset, for example, consider M_{ui} = {4,4,4.5,2} be user ratings for (Toy Story movie), w = [0.22,0.44, 0.76, 0.40] and b = [0.33] is a random weight and bias for our model, a predicted linear-model for the first class is:

$$MLc_1\, 0.22*4.0+0.44*4.0+0.76*4.5+0.40*2+0.33MLc_1 = 7.19$$

Now, let's assume that the ML result for the other class is as ([8.55,5.40,8.5,10.09,6.59]). The result of the log score (linear predictor) will then be translated to probability value using the SoftMax approach. The formula is shown in Eq. 5:

$$S(ci) = \frac{e^{c_i}}{\sum_{k=1}^{n} e^{c_i}} \tag{5}$$

where as, C_i is a class name. *S*, SoftMax function.

The SoftMax approach requires that all other class scores be included when calculating the probability of one of the classes, for instance:

$$S(c_1) = \frac{e^{7.19}}{e^{7.19} + e^{8.55} + e^{5.40} + e^{8.5} + e^{10.09} + e^{6.59}}$$

In the final phase of the models, the SoftMax output with target one-hot encoding will be fed into the cross-entropy method to produce a predicted class.

After using MLR to predict the supercluster class for a new active user, a modified Mahalanobis distance indicator was used to estimate the sub-cluster, assuming; kc_1 = {3,4,5,2}, M_{Ui} = {4,4,4.5,2} and, σ = {1,0.2,0,5,1,5}.

The Mahalanobis distance between them is as bellow:

$$MD(M_{ui}, kc_i) = \sqrt{\frac{(3-4)^2}{(1)^2} + \frac{(4-4)^2}{(0.2)^2} + \frac{(5-4.5)^2}{(0.5)^2} + \frac{(2-2)^2}{(1.5)^2}}$$

The new active user will be placed in the cluster with the closest distance. To discover the top N similar users, the Euclidean distance between the new active user and each user in the same cluster was computed. Based on that, the new user will be recommended to show movies with the highest rating with threshold-rating, as shown in Table 10.

Performance and Evaluation Metrics

A popular evaluation metric called the Silhouette score was utilized to verify the performance of clustering results with different values of k. The Silhouette score can be computed from Eq. (6).

$$S = (b-a) / \max(a,b) \tag{6}$$

Abd and Aminifar (2022b). the best score was obtained when the value of k was 6. As Table 11 depicts the outcome, the highest score was 0.822. visualizing the result of the silhouette score is in Figs. 6, 7, 8.

Finally, to evaluate the classification model we used classification report and mean absolute error metrics. The result shows 1.00 for precision and 0.99 for recall which indicates a high prediction accuracy result. A comparative review is shown in Table 12.

The best MEA result obtained was 0.04 and 0.05 for train and testing the MLR model using Mahalanobis similarity with Euclidean similarity measure. Figures 9 and 10 show the result.

Table 10: Example of recommended movies based on user rating

Movie Id	Name
1	Toy (1992)
2	What the #$*! Do We Know!? (a.k.a. What the Bleep Do We Know!?) (2004)
3	Pan's Labyrinth (Laberinto del fauno, El) (2006)
4	Kids (1995)
5	Demolition Man (1993)
6	Philadelphia (1993)
7	Mission: Impossible III (2006)
8	Indiana Jones and the Temple of Doom (1984)
9	Fugitive, The (1993)
10	Atlantis: The Lost Empire (2001

Table 11: The k-means Silhouette Score with a different number of features.

Value of k	Number of features = 2	Number of features = 3	Number of features = 4
2	0.210	0.20	0.10
4	0.410	0.30	0.13
6	0.822	0.52	0.43
8	0.270	0.21	0.19
10	0.520	0.41	0.40

Table 12: the summary of reviewed papers.

Authors and Publication Year	Technique(s)	Method	Contribution(s)	Result
Mustaqeem *et al.* (2020)	K-means-clustering and sub-clustering	Numerical cleaner filter	Solve missing values and outlier issues	Give accurate medical advice recommendations in the quickest time possible
Zhang *et al.* (2017)	Matrix factorization	Merging topic modeling and text sentiment analysis with a Hybrid Matrix Factorization model (HMF)	Find user ratings in their emotional review and enhance standard matrix factorization by identifying topic distributions of user choice	Give a more reliable prediction as well as a significantly more accurate recommendation
Wasid and Ali (2018)	K-means clustering	Accommodate multicriteria rating into traditional RS	Solve multidimensionality issues	Create a collection of neighborhoods that is more similar to target users
Sun and Dong (2017)	Enhanced k-mean clustering	Time impact factor matric	Devise a new way of monitoring items that are rated several times	Predict user interest level indirectly
Natarajan *et al.* (2020)	Enhanced Matrix factorization	Matrix factorization	Address the data sparsity and cold start issue	Improve recommendation accuracy
Ahuja *et al.* (2019)	K-Nearest neighbor and K-means clustering	WCSS method with the elbow approach	Focus on introducing several ML and RS methods	Make better recommendations based on RMSE value
Yassine *et al.* (2021)	K-means clustering	PCA, SVD and user demographic	Reducing dataset 'dimensionality'	Enhance performance and reduce response time
Hu *et al.* (2020)	Item-based CF	’’Multiplex implicit feedbacks’’, PCC and VSS	Solving the issues of data sparsity	Improved RS accuracy
Thakkar *et al.* (2019)	User-based and Item-based CF	Multiple-linear regression	Decreasing prediction error	----
Chen *et al.* (2021)	User-based CF	“Users positive and negative profile”	Address the unrated items	Enhance the MAX index for standard CF
Pradhan *et al.* (2021)	CF, CB K-NN	Hybrid Movie RS	Handling the CF drawbacks	Manage massive datasets efficiently
Panchal *et al.* (2022)	for Item-based CF Naïve-Bias for sentiment-analysis	Sentiment analyses	Lowering the dimensional data’s noise	Improve accuracy, faster execution
Gupta *et al.* (2020)	K-NN CF	---	Handling the CF drawbacks	Boost the reliability, and accuracy as well as efficiency of RS

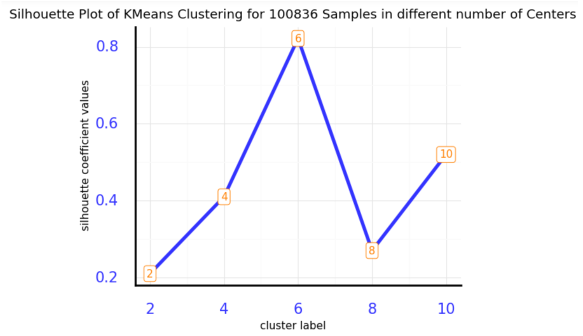


Fig. 6: The relationship between the value of k and silhouette score when the number of features = 2

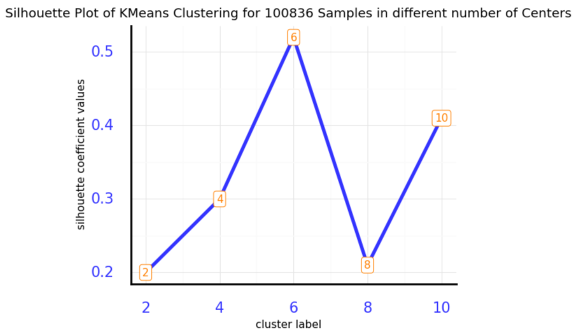


Fig. 7: The relationship between the value of k and silhouette score when the number of features = 3

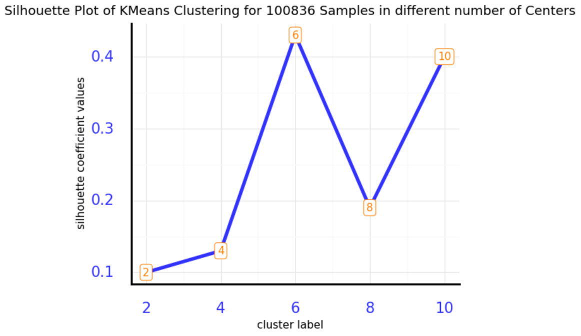


Fig. 8: The relationship between the value of k and silhouette score when the number of features = 4

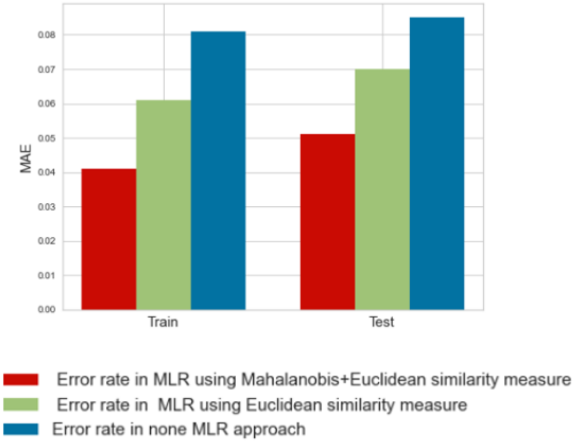
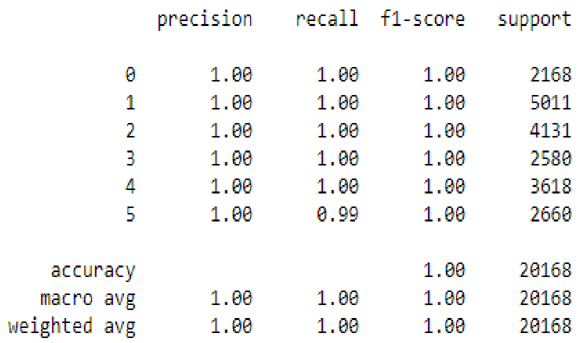


Fig. 9: Comparison graph of MAE with existing RS techniques



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2168
           1       1.00      1.00      1.00      5011
           2       1.00      1.00      1.00      4131
           3       1.00      1.00      1.00      2580
           4       1.00      1.00      1.00      3618
           5       1.00      0.99      1.00      2660

    accuracy                           1.00     20168
   macro avg       1.00      1.00      1.00     20168
weighted avg       1.00      1.00      1.00     20168
```

Fig. 10: Precision, recall, and f1-score result for testing the MLR model

Discussion and Comparison of Similar Works

Several researchers have worked to enhance RS in terms of accuracy, processing speed, and complexity. However, according to publications we evaluated, their approaches performed better in accuracy, response time, etc.

In 2020 research on medical advice recommendations, the researchers worked on reducing search space using k-means as a sub-clustering technique; the result shows that using sub-clustering fulfills some of the performance criteria, such as response time and accurate results. In 2021, the researcher used PCA before clustering the Dataset; their result demonstrates that using PCA reduced the dimensionality problem and response time. However, many researchers demanded to improve just one area of RS and they did not use a probabilistic in clustering user ratings which enhance the accuracy of recommendation. In this proposed approach, we focused on how to recommend items that users deserve with high accuracy using a powerful probabilistic model.

Conclusion

Multicriteria ratings, as opposed to single criteria ratings, can aid in more realistically describing customers' favorite items. It is possible to indicate what users deserve when they rate an object based on multiple factors. However, multicriteria ratings cause multidimensionality problems during the process. In this study, we have introduced a novel CFRS strategy to solve data processing complexity and accuracy issue in CFRS due to normalizing the Dataset before the feature reduction technique. Our proposed approach achieved significant enhancement in the clustering phase regarding inertia value (distance between each point and its centroid). As a result, the applied PCA dataset had a lower inertia value in contrast to an original dataset with the same number of clusters with (125107.546 and 4906.362) values for the original and extracted features Dataset, respectively.

Further, due to the sub-clustering technique, which reduced the search space, the proposed approach performs better than similar work in terms of faster similarity discovery.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Zhila Yaseen Taha: Review, Research, Analyze, and Simulation.

Sadegh Aminifar: Supervision and consultation during reviewing, result, and discussion

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abd, M. H. M., & Aminifar, S. (2022a). A Demodulator Selection Model for Received FSK and ASK Signals. Neuro Quantology, 20(10), 2181-2186. https://doi.org/10.14704/nq.2022.20.10.NQ55188

Abd, M. H. M., & Aminifar, S. (2022b). Intelligent Digital Signal Modulation Recognition using Machine Learning.

Ahuja, R., Solanki, A., & Nayyar, A. (2019, January). A movie recommender system using K-Means clustering and K-Nearest Neighbor. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 263-268). IEEE. https://ieeexplore.ieee.org/abstract/document/8776969

Ajaegbu, C. (2021). An optimized item-based collaborative filtering algorithm. *Journal of Ambient Intelligence and Humanized Computing*, *12*(12), 10629-10636. https://link.springer.com/article/10.1007/s12652-020-02876-1

Aminifar, S., Khoei, A., Haidi, K., & Yosefi, G. (2006). A digital CMOS fuzzy logic controller chip using new fuzzifier and max circuit. AEU-International Journal of Electronics and Communications, 60(8), 557-566. https://doi.org/10.1016/j.aeue.2005.11.003

Aminifar, S. (2020). Uncertainty Avoider Interval Type II Defuzzification Method. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/5812163

Aminifar, S., & Marzuki, A. (2013a). Horizontal and vertical rule bases method in fuzzy controllers. *Mathematical Problems in Engineering*, *2013*. https://doi.org/10.1155/2013/532046

Aminifar, S., & Marzuki, A. (2013b). Uncertainty in interval type-2 fuzzy systems. *Mathematical Problems in Engineering*, *2013*. https://doi.org/10.1155/2013/452780

Brownlee, J. (2021). Stacking Ensemble Machine Learning with Python. *Machine Learning Mastery. https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/*

Chen, Y. L., Yeh, Y. H., & Ma, M. R. (2021). A movie recommendation method based on users' positive and negative profiles. *Information Processing & Management*, *58*(3), 102531. https://doi.org/10.1016/j.ipm.2021.102531

Gupta, M., Thakkar, A., Gupta, V., & Rathore, D. P. S. (2020, July). Movie recommender system using collaborative filtering. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 415-420). IEEE. https://ieeexplore.ieee.org/abstract/document/9155879

Hamzah, N. A., Kek, S. L., & Saharan, S. (2017). The Performance of K-Means and K-Modes Clustering to Identify Cluster in Numerical Data. *Journal of Science and Technology*, *9*(3). https://publisher.uthm.edu.my/ojs/index.php/JST/article/view/2038

Hu, Y., Xiong, F., Lu, D., Wang, X., Xiong, X., & Chen, H. (2020). Movie collaborative filtering with multiplex implicit feedbacks. *Neurocomputing*, *398,* 485-494. https://doi.org/10.1016/j.neucom.2019.03.098

Jader, R. and Aminifar, S., (2022a). Fast and accurate artificial neural network model for diabetes recognition. *Neuro a Quantology*, 20(10), pp, 2187-2196.

Jader, R., & Aminifar, S. (2022b). Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach. *Applied Computational Intelligence and Soft Computing*, *2022*. https://doi.org/10.1155/2022/974957

Jader, R. F., Aminifar, S., & Abd, M. H. M. (2022). Diabetes detection system by mixing supervised and unsupervised algorithms. *Journal of Studies in Science and Engineering*, *2*(3), 52-65. http://www.engiscience.com/index.php/josse/article/view/josse2022234

Kant, S., Mahara, T., Jain, V. K., Jain, D. K., & Sangaiah, A. K. (2018). LeaderRank based k-means clustering initialization method for collaborative filtering. *Computers & Electrical Engineering*, *69*, 598-609. https://doi.org/10.1016/j.compeleceng.2017.12.001

Kuo, R. J., Zheng, Y. R., & Nguyen, T. P. Q. (2021). Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering. *Information Sciences*, *557*, 1-15. https://doi.org/10.1016/j.ins.2020.12.051

Liu, Y., Nie, J., Xu, L., Chen, Y., & Xu, B. (2017, September). Clothing recommendation system based on advanced user-based collaborative filtering algorithm. In *International Conference on Signal and Information Processing, Networking and Computers* (pp. 436-443). Springer, Singapore. https://link.springer.com/chapter/10.1007/978-981-10-7521-6_53

Marjai, P., Lehotay-Kéry, P., & Kiss, A. (2021). Document similarity for error prediction. *Journal of Information and Telecommunication*, *5*(4), 407-420. https://doi.org/10.1080/24751839.2021.1893496

Marzuki, A., Tee, S. Y., & Aminifar, S. (2014). Study of fuzzy systems with Sugeno and Mamdanitype fuzzy inference systems for determination of heartbeat cases on Electrocardiogram (ECG) signals. *International Journal of Biomedical Engineering and Technology*, 14(3), 243-276.

Mustaqeem, A., Anwar, S. M., & Majid, M. (2020). A modular cluster based collaborative recommender system for cardiac patients. *Artificial Intelligence in Medicine*, *102*, 101761. https://doi.org/10.1016/j.artmed.2019.101761

Natarajan, S., Vairavasundaram, S., Natarajan, S., & Gandomi, A. H. (2020). Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data. *Expert Systems with Applications*, *149*, 113248. https://doi.org/10.1016/j.eswa.2020.113248

Panchal, B. Y., Dave, K., Darji, H., Husain Bohara, M., & Talati, B. (2022). An Effective Movie Recommendation System Using Collaborative Filtering and User Review Sentimental Analysis. *SSRN 4151680*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4151680

Pradhan, R., Swami, A. C., Saxena, A., & Rajpoot, V. (2021, March). A Study on Movie Recommendations using Collaborative Filtering. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1119, No. 1, p. 012018). IOP Publishing. https://iopscience.iop.org/article/10.1088/1757-899X/1119/1/012018/meta

Qian, Y., Zhang, Y., Ma, X., Yu, H., & Peng, L. (2019). EARS: Emotion-aware recommender system based on hybrid information fusion. *Information Fusion*, *46*, 141-146. https://doi.org/10.1016/j.inffus.2018.06.004

Sun, B., & Dong, L. (2017). Dynamic model adaptive to user interest drift based on cluster and nearest neighbors. *IEEE Access*, *5*, 1682-1691. https://ieeexplore.ieee.org/abstract/document/7864348

Thakkar, P., Varma, K., Ukani, V., Mankad, S., & Tanwar, S. (2019). Combining user-based and item-based collaborative filtering using machine learning. In *Information and Communication Technology for Intelligent Systems* (pp. 173-180). Springer, Singapore. https://link.springer.com/chapter/10.1007/978-981-13-1747-7_17

Wasid, M., & Ali, R. (2018). An improved recommender system based on multi-criteria clustering approach. *Procedia Computer Science*, *131*, 93-101. https://doi.org/10.1016/j.procs.2018.04.190

Xue, F., He, X., Wang, X., Xu, J., Liu, K., & Hong, R. (2019). Deep item-based collaborative filtering for top-n recommendation. *ACM Transactions on Information Systems (TOIS)*, *37*(3), 1-25. https://doi.org/10.1145/3314578

Yassine, A. F. O. U. D. I., Mohamed, L. A. Z. A. A. R., & Al Achhab, M. (2021). Intelligent recommender system based on unsupervised machine learning and demographic attributes. *Simulation Modelling Practice and Theory*, *107*, 102198. https://doi.org/10.1016/j.simpat.2020.102198

Zhang, Y., Chen, M., Huang, D., Wu, D., & Li, Y. (2017). iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, *66*, 30-35. https://doi.org/10.1016/j.future.2015.12.001