Applied Information Technology And Computer Science Vol. 4 No. 1 (2023) 290-307

AITCS

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

Research on the Demographic Filtering Machine Learning in Movie Recommendation System

Lee Jia Yin[1], Noor Zuraidin Mohd Safar[1]*

[1]Fakulti Sains Komputer dan Teknologi Maklumat,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/aitcs.2023.04.01.018
Received 30 October 2022; Accepted 26 September 2022; Available online 30 June 2023

Abstract: In this era of big data explosion, movie recommendation system is widely used as an information tool by human. There are two common issues found in the machine learning movie recommendation system still undeniable: first, cold start, and second, data sparsity. To minimize the issues, a research study is conducted with objective to analyze, study, and test a decision-making algorithm that can solve cold start problem in a movie recommendation system with precise parameter. It involves the implementation of the proposed demographics filtering technique with k-means clustering method using tools like MATLAB, Microsoft Excel, and Microsoft Azure Machine Learning Studio. The research findings shall present the effects of the proposed demographic filtering for movie recommendations. Demographic filtering can group users into clusters based on parameter like gender, age group, and occupation. This research paper shall contribute demographic filtering studies as an alternative solution for the future work of technical development.

Keywords: Machine Learning, Recommendation System, Demographic Filtering, K-Means Clustering

1. Introduction

The advancement in technology is reaching new height every day and it is obviously noticeable. To manage a large amount of data, machine learning that builds analytical model automatically is introduced. People use machine learning to generate a recommender system that predicts most related recommendation using various computational statistics of datasets on the internet [1]. In fact, the recommender system is a kind of information filtering system that works predicting the preference or rating of an item. Generally, there are three recommendation system approaches. They are the content-based filtering, collaborative filtering, and hybrid filtering approach [2].

Interestingly, famous movie streaming service application like Netflix uses CineMatch as their proprietary recommender system since 2000. CineMatch is a software that embedded in the Netflix web

*Corresponding author: zuraidin@uthm.edu.my

site where it applies machine learning and data mining to analyze customer's preferences on movie-selection and recommend other movies that the customer might most likely enjoy [3].

Meanwhile, the movie recommendation system is also used and introduced widely with the rise of machine learning and recommendation system in various areas. The content-based movie recommendation system works based on the similarity of movie types and attributes. While the collaborative filtering movie recommendation system works based on past interactions between users and movies in the certain platform [4].

Generally, there are three main issues on the machine learning implementation to the existing recommendation system. Initially, the cold start problem can be found often in the situation of the new user starts using the recommendation system without any rating records. The data sparsity problem also occurs when the user has rated very few movies in the movie streaming application where it causes the recommendations given are not as accurate as expected. Furthermore, the prediction of the movie recommendation system can be not accurate if the parameter of a movie recommendation system is not clear and transparent.

Though we cannot deny that the movie recommendation application had brought us a lot of accessibilities in decision-making. The main purpose of this paper is to analyze, study and test an algorithm that can solve cold start problem in a movie recommendation system with precise parameter. The research result is expected to provide a technique that can work the best to solve several common challenges found in a machine learning movie recommendation system.

2. Related Work

This chapter studies the related work for the implementation of machine learning in movie recommendation system. It includes the literature study on the topic of recommendation system and machine learning. Furthermore, the literature review for demographics filtering is included in this chapter as well.

2.1 Recommendation system

Recommendation system is an information tool that helps users to get their desired items among a few choices. A recommendation system usually has an objective to predict the preference of a specific user upon an item or selection. User can find the best fit solution even if there are many items among the available choices. Recommendation system has been widely applied by many companies like Netflix, YouTube, Amazon, and others as it provides better service to their user and increases the company profit consequently.

Popular platform like Netflix uses recommendation system to suggest movies, Amazon uses recommendation system to suggest products, Spotify uses recommendation system to suggest songs, LinkedIn uses recommendation system to suggest jobs. All these essentials involve recommendation system. With the assistance of recommendation system, users can easily get what they want. It is difficult to develop a good recommendation system as user's preference keeps changing from time to time [5].

2.2 Machine learning

Machine learning falls in the category of artificial intelligence. Different from human learning, machine learning is a process of learning from experience which relies on data contrary. This makes the computers to learn on their own. Computers will keep modifying their actions to improve the actions to achieve more accuracy, which means more correct results will be obtained. With machine learning, many problems that require learning can be solved by the machine. There are six processes in a generic machine learning model and the processes are collection and preparation of data, feature selection, choice of algorithm, models and parameters selection, training, and performance evaluation [6].

Machine learning can be categorized into two main classes. The supervised learning and the unsupervised learning [7].

2.3 Demographic filtering

Demographics filtering is a kind of technique that uses demographic information to generate the most likely recommendation. Assume there are users sharing the same gender, age, occupation, etc., they are likely to have similar preference and favorite movies. Thus, users with similar demographics will be suggested with same movie recommendations in a demographics filtering movie recommendation system. Essentially, users will have specified data in their profile of movie watching platform or even movie recommendation system. The similarity or difference among the data of users will be performed to discover the laps in interests between users. This technique requires less complex computation as compared to the collaborative filtering and content-based filtering in recommendation system.

However, it requires a library data which contains a large amount of data for movies, ratings, and demographics from many users. In fact, demographics filtering is sometimes called as the enhanced collaborative filtering. Each movie is assigned with one or more genres in classes and the user is attracted to movies from the corresponding class. Generally, a demographics filtering recommendation system will involve user demographic information such as gender, age, occupation, etc. to perform group clustering [8].

Demographic filtering technique has strength as it is not based on user-item ratings, therefore it generates recommendation before user made any prior rating. This contributes to solve common problems like cold start and data sparsity that often observed in a machine learning movie recommendation system [9].

However, demographic filtering has its own disadvantage as well. As gathering the demographic data from system users, it may lead to privacy issues [9].

3. Methodology

This chapter studies about the methods and procedures used for the research on machine learning in movie recommendation system. There are sub-sections to cover the resources used, research framework, data selection, data pre-processing, parameter selection, k-means clustering and experimental setting for this project.

3.1 Resources used

The tools used for data analyzing and experimental setup purposes in this research study are listed as below:

- MATLAB
- Microsoft Excel PivotTable and PivotChart report
- Microsoft Azure Machine Learning Studio

MATLAB is a platform that is designed for high-level programming language, the MATLAB language, and acts as a numeric computing environment launched by MathWorks. It offers extensions to the desktop, support for parallel computing and GPU, and the “Live Editor”, which works to merge programs, descriptive text, output, and graphics into a formatted document [10]. In this research, MATLAB is utilized to perform data pre-processing, clustering chart plotting and data visualization for large file like ratings.dat.

Microsoft Excel is a spreadsheet software program that is useful and powerful for data visualization and analysis. It allows users to format, organize and calculate data in a spreadsheet. This eases the information easier to view as data is modified or changed. The PivotTable and PivotChart report features from Microsoft Excel is applied in analyzing and visualizing data from files like movies.dat and users.dat in this research study.

Microsoft Azure Machine Learning Studio is the central point of contact for machine learning computation in the Azure cloud. Since the Microsoft Azure Machine Learning Studio is one type of platform as a service (PaaS), it is hosted on virtual machines (VMs) and does not have a standard system requirement. In this research study, Microsoft Azure Machine Learning Studio is utilized as an experiment environment for machine learning model training, deployment, result evaluation for analysis purpose as well.

3.2 Research framework

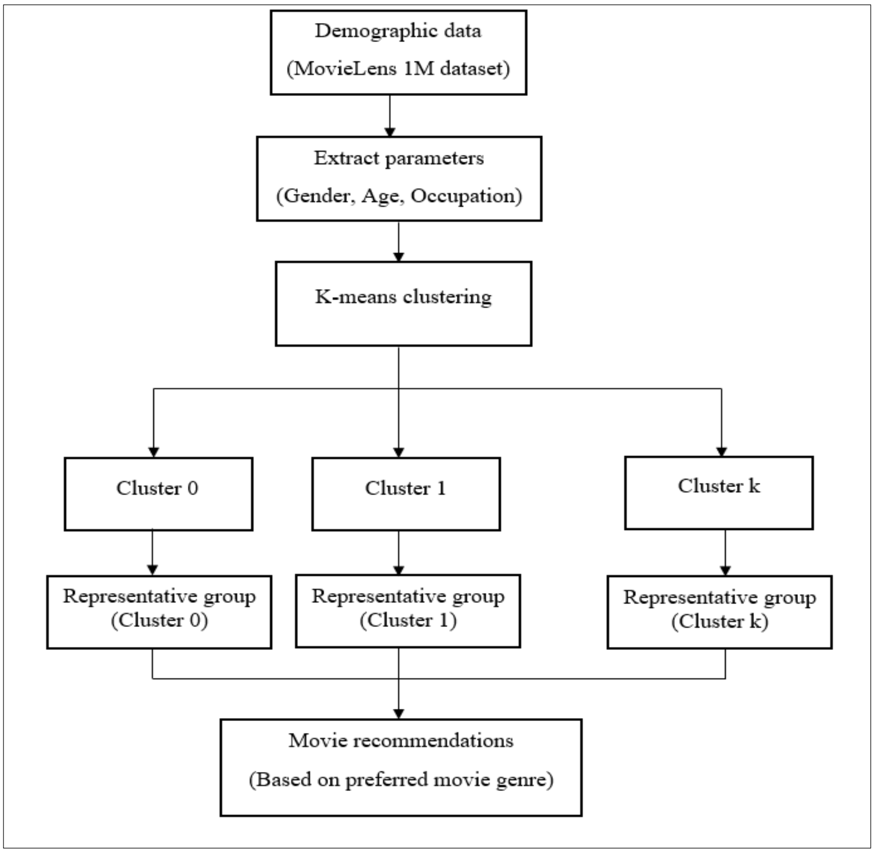


Figure 1 : Framework of proposed demographic filtering movie recommendation system

The research framework of demographic filtering movie recommendation system proposed in this paper is shown as in Figure 1.

The research framework shows that the demographic data will be the independent variable and the movie recommendations will be the dependent variable. In other words, the movie recommendation result will depend on the user demographic data. Besides, the K-means clustering is used in the research study as the mediator variable that indicates the effects of demographic data on the movie recommendations. K-means clustering is known as partitioning algorithm as well. K-means clustering is used in this research study as it is an easy-to-implement algorithm that can perform fast clustering for a large group of data variables [11]. After the data is scattered into clusters based on the demographic similarities, a representative group will be introduced along with the preferred movie genre for that cluster. Movie recommendations will be made based on what cluster the user been placed.

3.3 Data selection

The dataset involved in this research project is the MovieLens 1M. The dataset is a part of GroupLens Research Project from the University of Minnesota. The MovieLens 1M Dataset is imported, where it consists of 1000209 ratings of approximately 3900 movies collected from 6040 users. The demographic information in the file of users packaged in the MovieLens 1M Dataset is the main reason to explain why it is used as compared to other datasets offered by MovieLens. After all, it is publicly available [12].

3.4 Data pre-processing

Data pre-processing is one of the most crucial processes in the training of a machine learning model. This process happens before processing a collection of data or dataset into a model. The purpose of data pre-processing is to ensure the accuracy, efficiency, and meaning of the analysis on the research model. Besides, data pre-processing can bring consequences to increase the transparency level of the machine learning process by analyzing each specific step, thus introducing better and fairer model for further improvement [13]. In this research study, data pre-processing is done on MATLAB.

In the data pre-processing stage of this research study, function like ismissing() is called to find the number of missing values in the dataset imported. Next, rmmissing() is called to remove the missing entries in the dataset imported. Furthermore, removevars() is called to remove unnecessary column in the dataset imported from MovieLens 1M dataset, for example, the zip code variable in the users.dat file as it is not the parameter in this research study.

3.5 Parameter selection

Parameter is a numerical quantity or attribute used to describe the characteristic of population in research. To implement the proposed demographic filtering methodology, there are a few parameters to be considered for measurement from the files of MovieLens 1M dataset. The parameters measured and involved are the user gender, age, and occupation, followed by movie genres. The parameters are explained as shown in Table 1, Table 2, Table 3, and Table 4. There are three tables describe about the demographic's attributes of user dataset and one table describes about the genres of movie shown as below.

Table 1: Gender

ID	Attribute
M	Male
F	Female

Table 2: Age

ID	Attribute
1	Under 18
18	18 – 24
25	25 – 34
35	35 – 44
45	45 – 49
50	50 – 55

56	Above 55

Table 3: Occupation

ID	Attribute
0	Other/Not specified
1	Academic/Educator
2	Artist
3	Clerical/Admin
4	College/Grad student
5	Customer service
6	Doctor/Health care
7	Executive/Managerial
8	Farmer
9	Homemaker
10	K-12 student
11	Lawyer
12	Programmer
13	Retired
14	Sales/Marketing
15	Scientist
16	Self-employed
17	Technician/Engineer
18	Tradesman/Craftsman
19	Unemployed
20	Writer

Table 4: Genres

Attribute
Action
Adventure
Animation

Children’s
Comedy

Table 4: (cont)

Crime
Documentary
Drama
Fantasy
Film-Noir
Horror
Musical
Mystery
Romance
Sci-Fi

3.6 K-means clustering

The K-means clustering training is setup on Microsoft Azure Machine Learning Studio where it is a portal introduced by Microsoft Azure to conduct machine learning experiments. Microsoft Azure Machine Learning Studio introduces features to import dataset and add components to conduct machine learning experiments designed in the sense of researcher’s desire using pipeline.

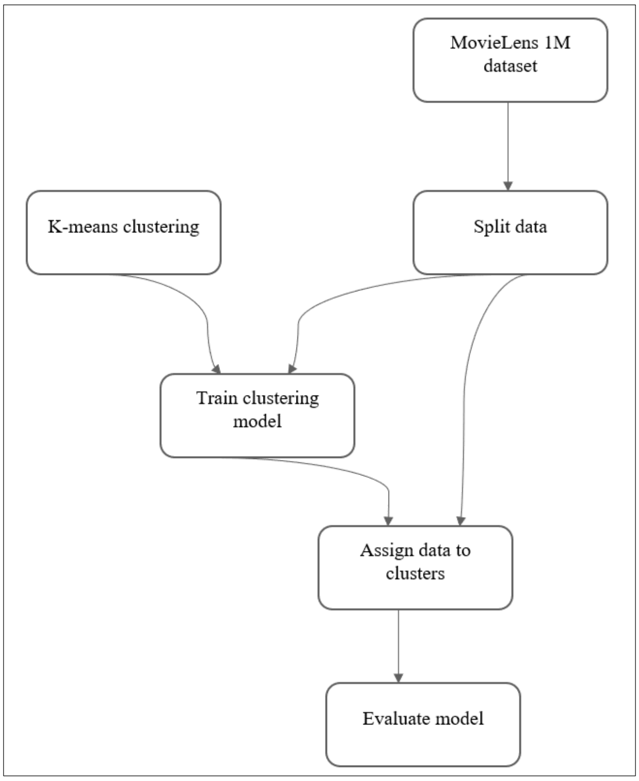


Figure 2 : Average distance to centroid testing with different number of clusters

In this research study, pipeline is used and designed following standard machine learning model training procedures as shown in Figure 2. The experiment design has an aim to train a clustering model on the user demographic information that comes from MovieLens 1M dataset. To conduct a successful machine learning model training experiment, the procedures must be strictly followed.

In this experiment, pipeline connects every essential procedure to deploy a complete machine learning model training experiment. The first step in the experiment design is to import the dataset. Since the dataset used in this experiment is the preprocessed MovieLens 1M dataset, procedures like selecting columns in dataset, clean missing data, and normalize data are already completed on MATLAB. The second step in the experiment is to split data. It is carried out to perform data splitting where 70 percent of the dataset will be used to train the clustering model and the remaining 30 percent of the dataset will be assigned into the clusters for evaluation purpose. Data splitting is an important step to avoid overfitting, as the overfitting can cause a machine learning model fits its training data too well and fails to reliably fit additional data that is out of training dataset. The third step is to link the 70 percent dataset into the clustering training model. At the same time, an untrained K-means clustering model is linked to train the clustering model. The next step is to assign the 30 percent riven data into the trained clustering model. The final step comes to evaluate the machine learning clustering model based on the demographic information from MovieLens 1M dataset.

The optimal number of clusters in K-means clustering model is tested by varying k from 2 to 10 within the similar experiment environment setup. Elbow curve method is applied to determine the number of centroid or cluster.

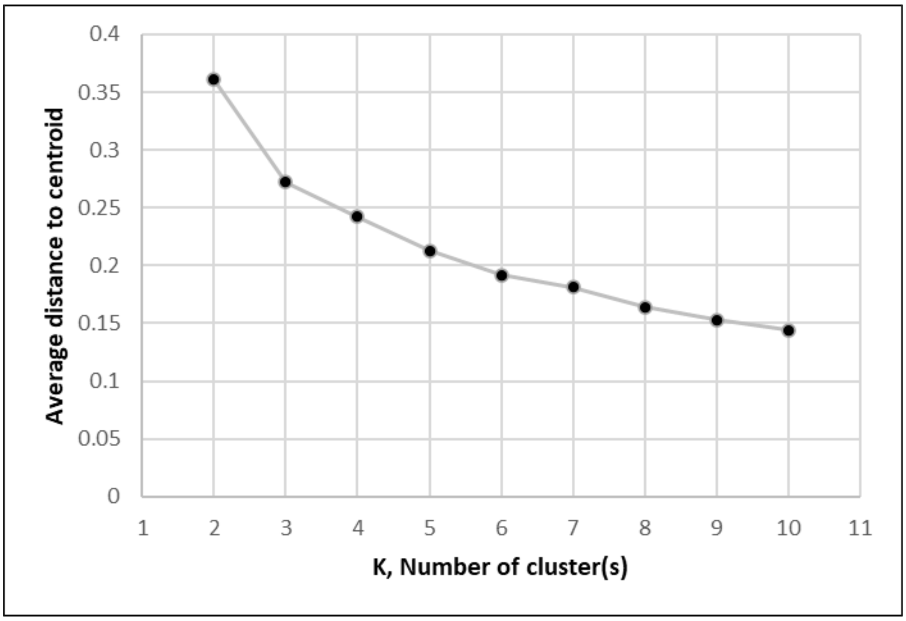


Figure 3 : Average distance to centroid testing with different number of clusters

Figure 3 shows the average distance to centroid testing with varying value of k for K-means clustering model. The result of testing is plotted and presented in the form of elbow curve as shown in Figure 3. To find the optimal number of clusters for training model based on the elbow curve method, the spot of elbow is visualized shall be taken. In this case, the spot falls at k equals to 3, where the average distance to cluster center falls suddenly at the testing on 3 clusters. Thus, it shows the optimal number of clusters or centroids to train the K-means clustering model is 3. The evaluation process shall show the result of clustering on demographic information.

3.7 Experimental setting

To consider the achievement of objective for this research study, all the experiments will be conducted using tools on MATLAB, Microsoft Excel, and Microsoft Azure Machine Learning Studio. The experiment setup and testing will be performed on a PC with an Intel(R) Core(TM) i5-9300H CPU at a speed of 2.40GHz and with 8GB RAM. The experiment activities are carried out with benchmark MovieLens 1M Dataset which contains files movies.dat, ratings.dat, and users.dat. Where movies.dat file contains information in the format of movie id, title, and genres. While ratings.dat file contains information in the format of user id, movie id, rating, and timestamp. And users.dat file contains information in the format of user id, gender, age, occupation, and zip code.

4. Results and Discussion

This chapter studies the result and data analysis of the research. The evaluation on the implementation of the proposed demographic filtering using the K-means clustering is reviewed and discussed. The analysis on parameter measured is included in this chapter as well.

4.1 Parameter analysis

There are a few important parameters being considered and tested from the user demographic and movie rating information to implement a demographic filtering movie recommendation system in this research study. The first parameter is the gender of user. From the dataset MovieLens 1M that is used in this research study, the gender of user is recorded and analyzed.

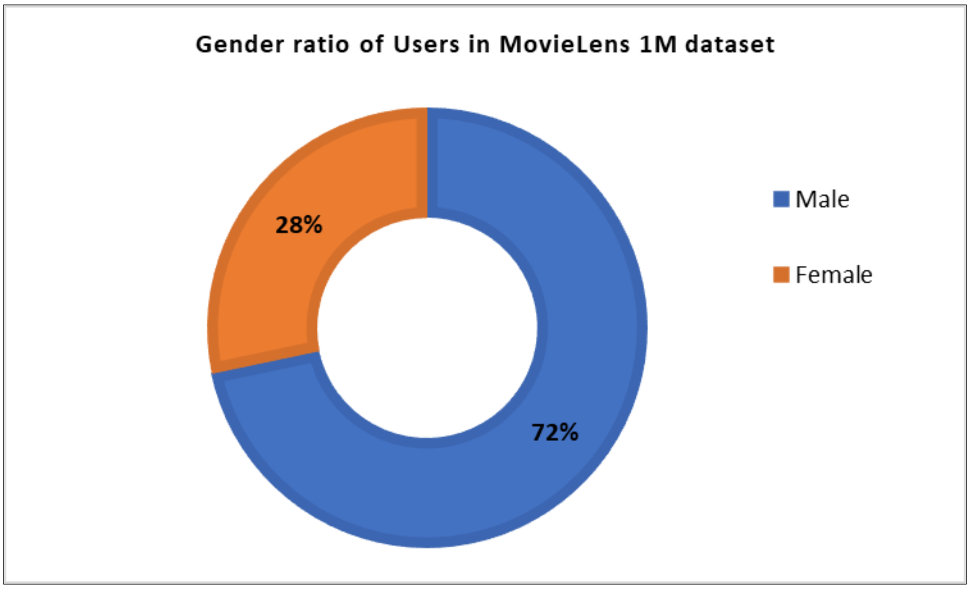


Figure 4 : Gender ratio of users in MovieLens 1M dataset

Figure 4 shows the gender ratio of users in MovieLens 1M dataset that is analyzed using Microsoft Excel's tool PivotTool and PivotChart report. From a total of 6040 counts of users, there are 4331 counts of male and 1709 counts of female contributions on the movie ratings in the MovieLens 1M dataset. There is more study of data being observed from the male users, where the gender ratio is 72:28 male to female in MovieLens 1M dataset.

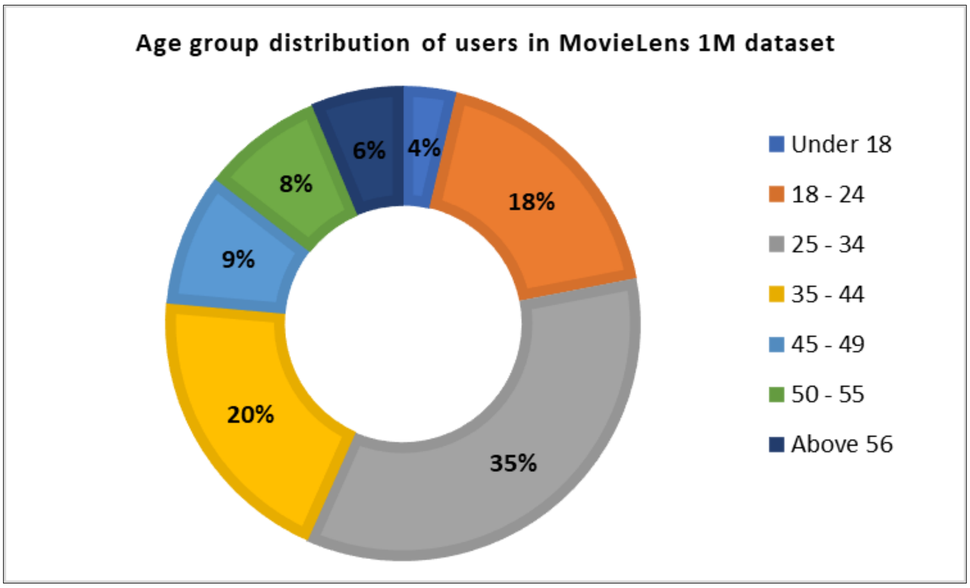


Figure 5 : Age group distribution of users in MovieLens 1M dataset

Figure 5 shows the age group distribution of users in MovieLens 1M dataset that is analyzed using Microsoft Excel's tool PivotTool and PivotChart report. From a total of 6040 counts of users, there are 222 counts of users in under 18 age group, 1103 counts of users in 18 to 24 age group, 2096 counts of users in 25 to 34 age group, 1193 counts of users in 35 to 44 age group, 550 counts of users in 45 to 49 age group, 496 counts of users in 50 to 55 age group, and 380 counts of users in above 56 age group in the MovieLens 1M dataset. There is more study of data being observed from the users in 25 to 34 age group where it shows a significant of domination which involves 35 percentage of users in MovieLens 1M dataset.

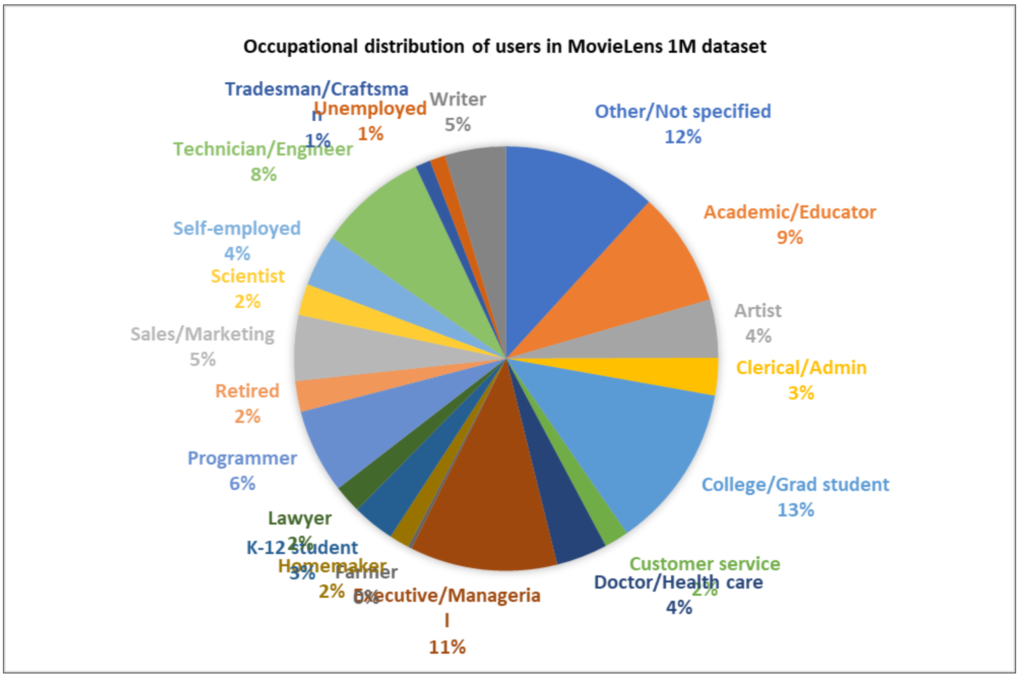


Figure 6 : Occupational distribution of users in MovieLens 1M dataset

Figure 6 shows the occupational distribution of users in MovieLens 1M dataset that is plotted using Microsoft Excel's tool PivotTool and PivotChart report. From a total of 6040 counts of users, there are 711 counts of users having other or not specified occupation, 528 counts of academics or educator, 267 counts of artist, 173 counts of clerical or admin, 759 counts of college or grad student, 112 counts of customer service, 236 counts of doctor or health care, 679 counts of executive or managerial, 17 counts of farmer, 92 counts of homemaker, 195 counts of k-12 student, 129 counts of lawyer, 388 counts of programmer, 142 counts of retired, 302 counts of sales or marketing, 144 counts of scientist, 241 counts of self-employed, 502 counts of technician or engineer, 70 counts of tradesman or craftsman, 72 counts of unemployed, and 281 counts of writer in the MovieLens 1M dataset. Figure 6 shows the occupational distribution of users from MovieLens 1M dataset which consists of 21 different occupations.

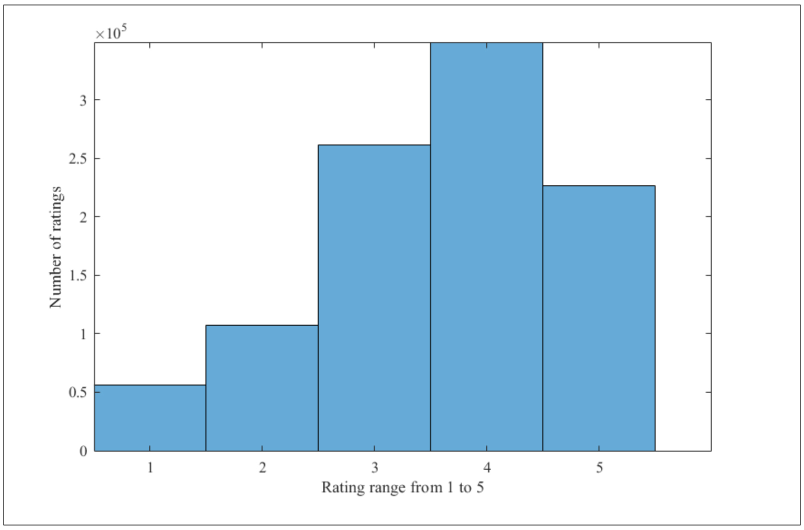


Figure 7 : The movie ratings distribution in MovieLens 1M dataset

Figure 7 presents the movie ratings distribution of MovieLens 1M dataset. The analysis process on movie ratings is computed by using the MATLAB's chart plotting tool. The count for movie rating of one is 56174, which is the least appointed movie rating. The count for movie rating of two is 107557.

The count for movie rating of three is 261197. The count for movie rating of four is 348971, which is the most appointed movie rating. The count movie rating of five is 226310. The total number of movie rating is 1000209 in MovieLens 1M dataset. The average of movie rating from MovieLens 1M dataset is 3.58 where it is left-skewed and negatively skewed.

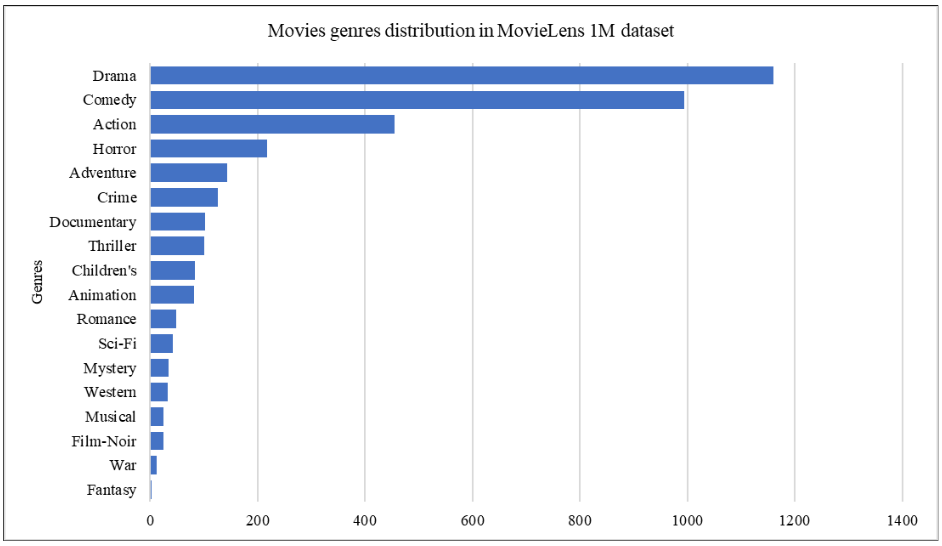


Figure 8 : The movie genres distribution in MovieLens 1M dataset

Figure 8 presents the genres distribution of movies in MovieLens 1M dataset that is visualized using Microsoft Excel's tool PivotTool and PivotChart report. From a total of 3686 counts of movies, there are 1161 counts of drama movie, 994 counts of comedy movie, 455 counts of action movie, 217 counts of horror movie, 143 counts of adventure movie, 126 counts of crime movie, 102 counts of documentary movie, 100 counts of thriller, 83 counts of children's movie, 82 counts of animation movie, 49 counts of romance, movie, 42 counts of sci-fi movie, 35 counts of mystery movie, 33 counts of western movie, 25 counts of musical movie, 25 counts of film-noir movie, 12 counts of war movie, and two counts of fantasy movie. Figure 4.5 shows the movie genres distribution in MovieLens 1M dataset, where most of the movies are tagged as drama and comedy genre.

4.2 Result evaluation

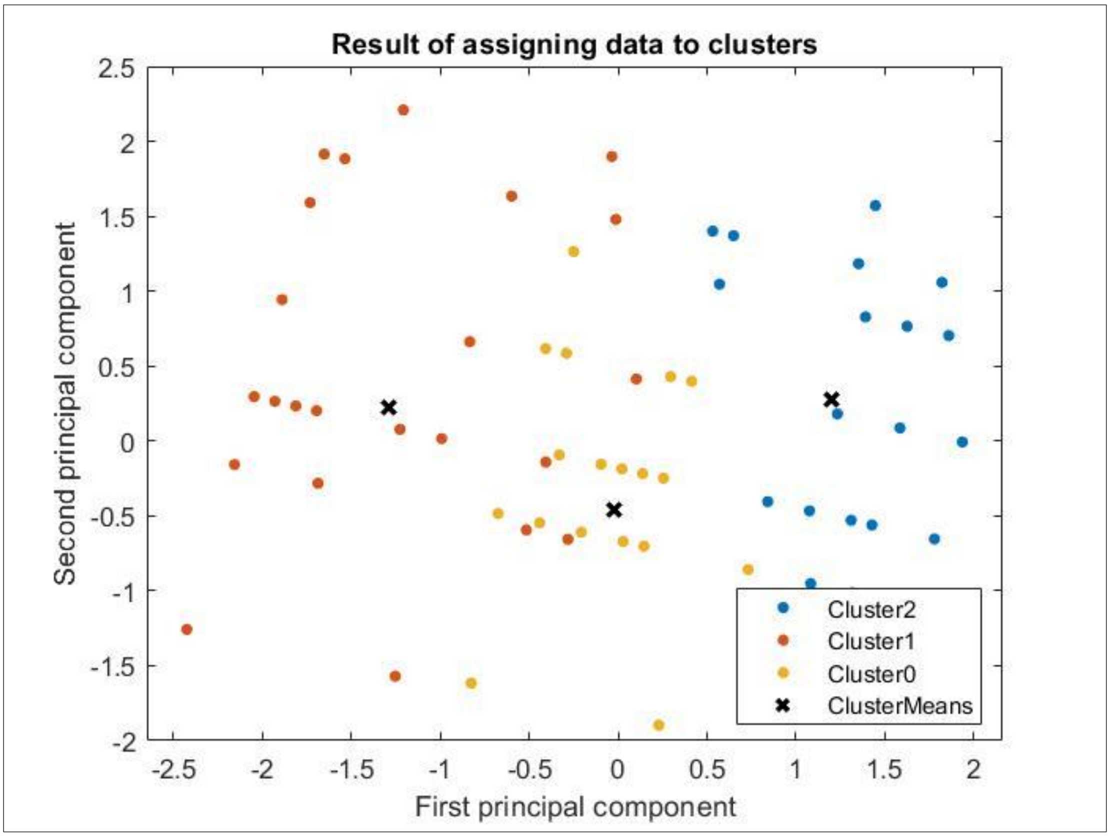


Figure 9 : Result of assigning data to clusters

Figure 9 shows the visualization of data assigned to clusters referred to the top 100 results of the experiment on Microsoft Azure Machine Learning Studio in this research study. The scatter plot in Figure 9 presents the result of assigning data to the trained clustering model in the experiment where the result is acceptable as the users are visibly divided into three different clusters based on their gender, age group, and occupation. The scatter plot is generated on MATLAB.

Table 5 below shows the clusters distribution representative group based on the top 100 results of the experiment on Microsoft Azure Machine Learning Studio in this research study. The user with the least distance to cluster center is chosen as the representative group in that cluster. Cluster 0 has a representative group from male, college or grad student, aged 25 to 34. Cluster 1 has a representative group from female, executive or managerial, aged 25 to 34. Cluster 2 has a representative group from male, sales or marketing aged 35 to 44. Table 5 below shows user from different cluster has various preferred movie genre where the preferred movie genre in Cluster 0 is action, adventure, comedy, drama, and war. Cluster 1 has preferred movie genre of comedy, crime, drama, horror, romance, and sci-fi. And Cluster 2 has preferred movie genre of action, comedy, drama, film-noir, mystery, and thriller.

Table 5: Clusters distribution

Cluster	Representative group	Distance to cluster center	Preferred movie genre
0	Male, 25, 4	0.091596	Action, Adventure, Comedy, Drama, War

1	Female, 25, 7	0.108244	Comedy, Crime, Drama, Horror, Romance, Sci-Fi
2	Male, 35, 14	0.083404	Action, Comedy, Drama, Film-Noir, Mystery, Thriller

4.3 Virtual experiment

For instance, if there is a new virtual User 1 arrives to the proposed demographic filtering machine learning movie recommendation system with information as shown as in Table 6 below.

Table 6: New virtual User 1 information

Demographic	Information
Gender	Male
Age group	18 – 24
Occupation	College/Grad student

Based on the finding of the assigning data to clusters experiment on Microsoft Azure Machine Learning Studio. The new virtual User 1 will be assigned to the Cluster 0 with the greater similarity in combination of gender, age group, and occupation. The movie recommendations will be made based on the preferred movie genre in his cluster assigned, which involves action, adventure, comedy, drama, and war in this case. The movie recommendations for virtual User 1 are shown as in Table 7 below, but not limited to them.

Table 7: Movie recommendations to virtual User 1

Movie	Genre
Knightriders (1981)	Action, Adventure, Drama
Missing in Action (1984)	Action, War
The Messenger: The Story of Joan of Arc (1999)	Drama, War
Bootmen (2000)	Comedy, Drama
Tigerland (2000)	Drama

The experiment findings should show the feasibility to support the implementation of demographic filtering approach in a machine learning movie recommendation system at primary. An effective clustering process based on user demographic information can bring a productive user categorization or grouping. And based on the clusters, ideally, users in same cluster will enjoy the recommended movie that come from a similar genre.

5. Conclusion

In the end, this research paper shall be a learning material supported by simple experiment setup to study the implementation of machine learning in movie recommendation system. The research study establishes the demographic filtering with K-means clustering algorithm to be the solution of cold start issue on a machine learning movie recommendation system. This paper shall contribute to the development of the recommendation system in movie industry, considering the performance of demographic filtering in a machine learning movie recommendation system. Nevertheless, the study of machine learning recommendation system can actually always be implemented in many other applications, but not limited to the movie industry. To succeed in the objectives of the research study in this project, demographics filtering is proposed with the complement of K-means clustering algorithm for the circumstances to solve common problems found in a machine learning movie recommendation system. Furthermore, the parameters that shall be involved in the demographic filtering machine learning movie recommendation system are studied and measured with visualization using tables and charts. Moreover, the MovieLens 1M dataset is chosen to test the clustering model for a movie recommendation system by assigning the data to clusters in the experiment. As the research paper is completed with the studies on machine learning in movie recommendation system, the future works should be considered for further research and development on applying machine learning in movie recommendation system. Considering demographic filtering is the proposed solution to solve common problems those exist in a general movie recommendation system, technical development to practically apply the demographic filtering machine learning model in a movie recommendation system should be studied and conducted in the future works. First, this future works shall generate a strong production that can prove the proposed demographic filtering is suitable to be implemented in a machine learning movie recommendation system. Second, the future works shall present a machine learning movie recommendation system that implements demographic filtering with further improvement to have a higher accuracy recommendation and involves less demographic information as parameter to protect the user privacy.

In this paper, demographic filtering is proposed to be implemented in movie recommendation system when compared to the traditional collaborative filtering, and content-based filtering. Demographic filtering is projected to solve cold start problem hence producing accurate movie recommendations for user in a movie recommendation system. After all, with an enduring exploitation in the future, there is still a lot of room for improvement on the implementation of artificial intelligence’s machine learning in human daily activities.

Acknowledgment

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support throughout this research study.

Appendix A

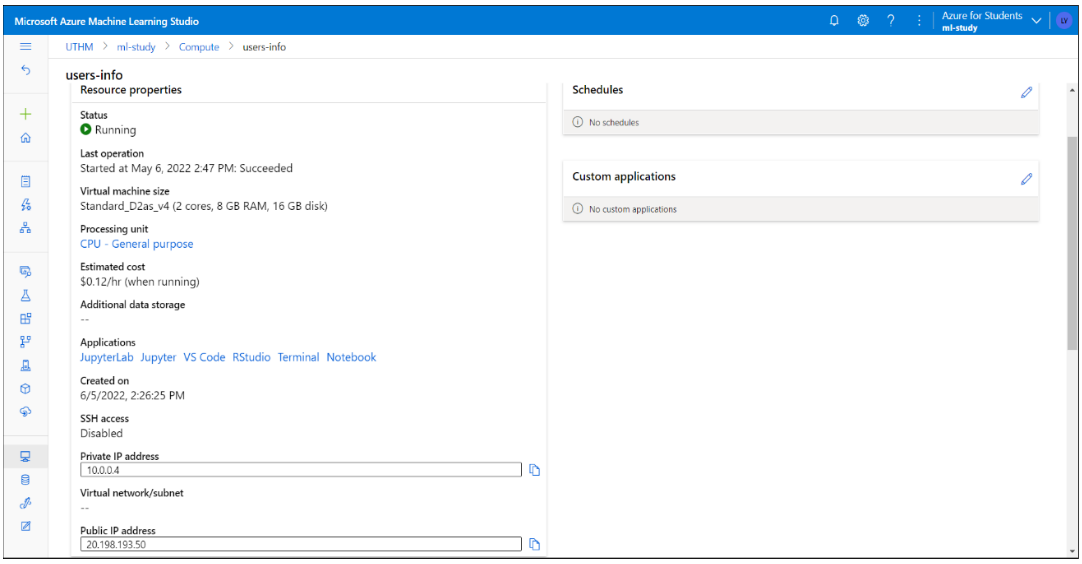


Figure 10 : Compute resource used on Microsoft Azure Machine Learning Studio

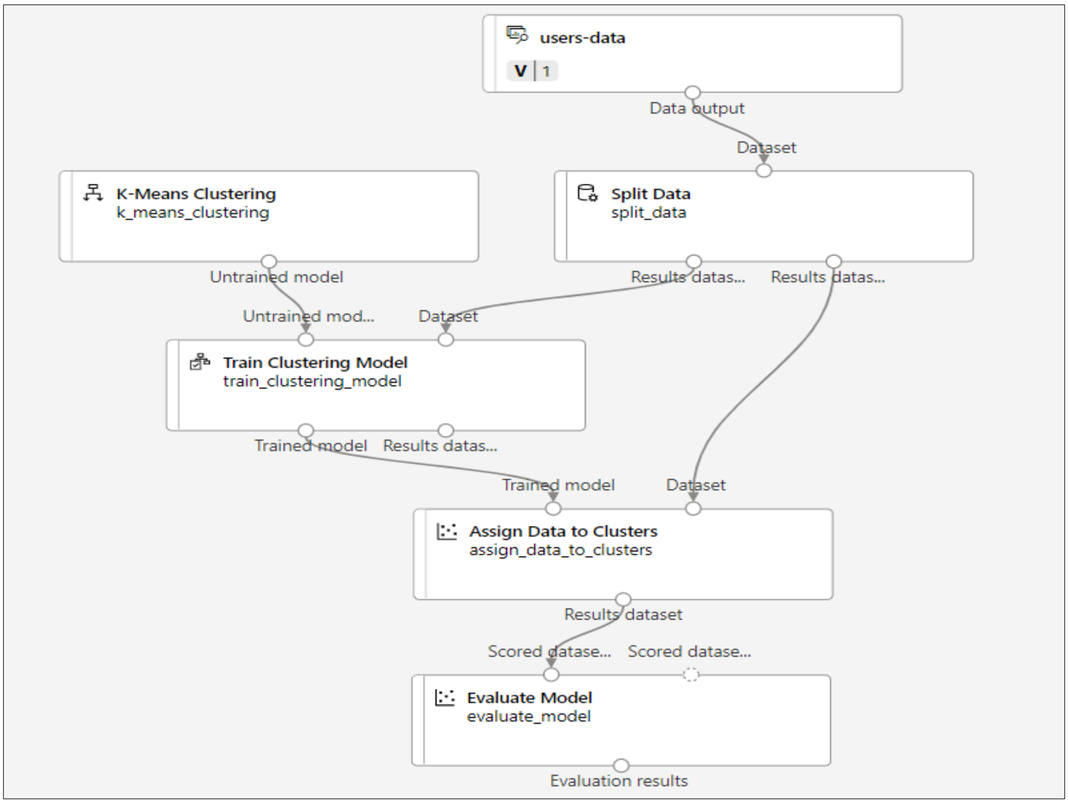


Figure 11 : K-means Clustering model training

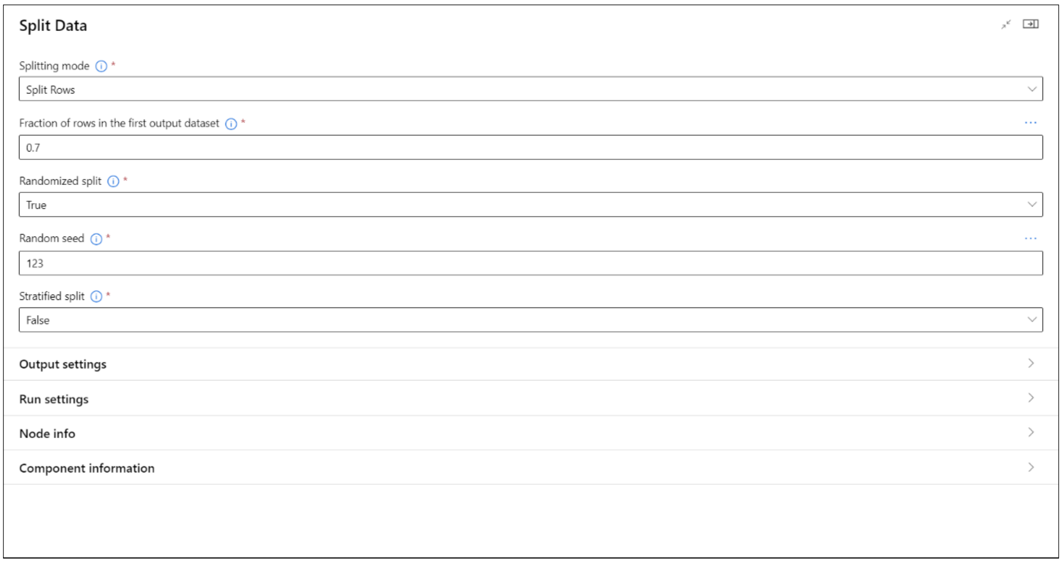


Figure 12 : Split data setup

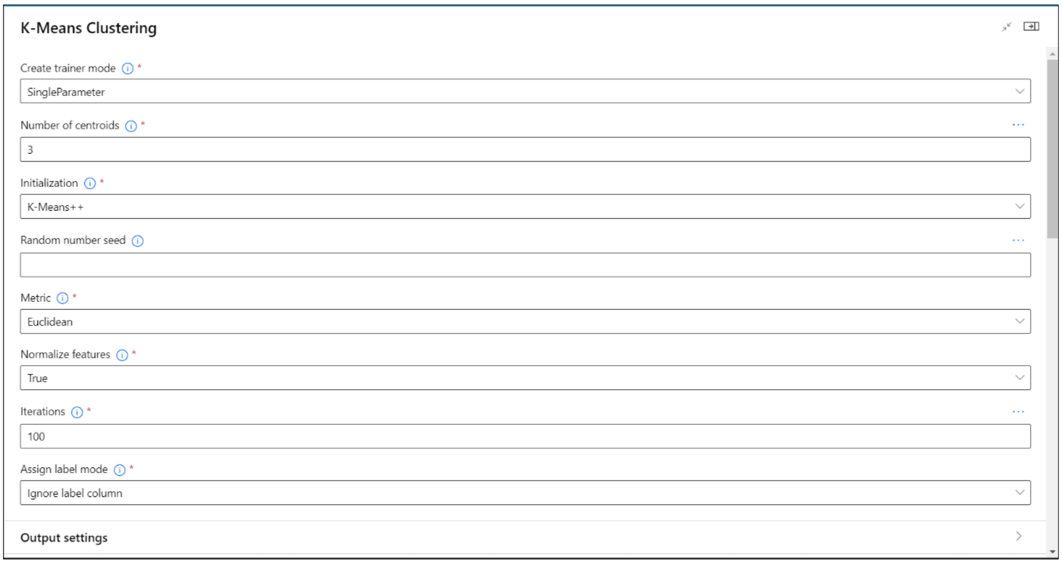
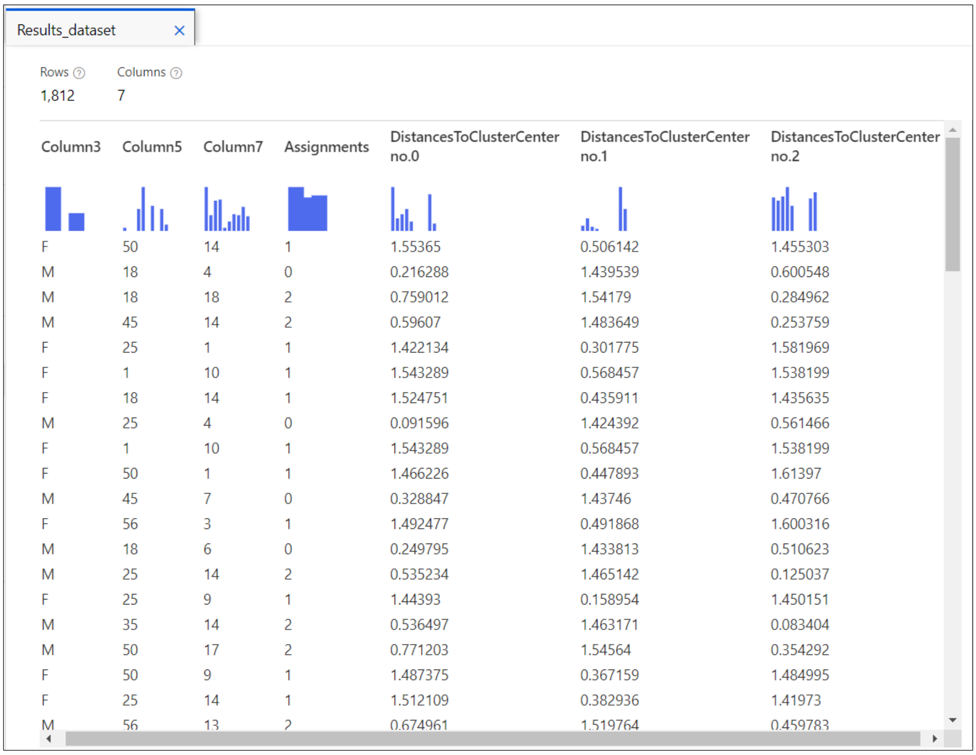


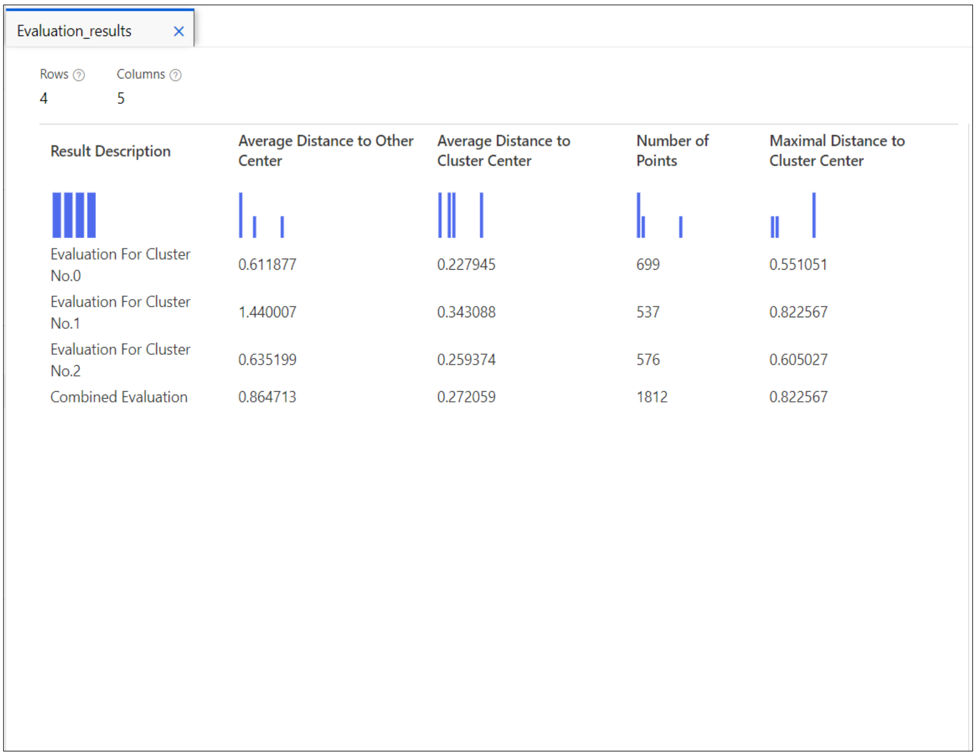
Figure 13 : K-means clustering setup


Results_dataset

Rows 1,812 Columns 7

Column3	Column5	Column7	Assignments	DistancesToClusterCenter no.0	DistancesToClusterCenter no.1	DistancesToClusterCenter no.2
F	50	14	1	1.55365	0.506142	1.455303
M	18	4	0	0.216288	1.439539	0.600548
M	18	18	2	0.759012	1.54179	0.284962
M	45	14	2	0.59607	1.483649	0.253759
F	25	1	1	1.422134	0.301775	1.581969
F	1	10	1	1.543289	0.568457	1.538199
F	18	14	1	1.524751	0.435911	1.435635
M	25	4	0	0.091596	1.424392	0.561466
F	1	10	1	1.543289	0.568457	1.538199
F	50	1	1	1.466226	0.447893	1.61397
M	45	7	0	0.328847	1.43746	0.470766
F	56	3	1	1.492477	0.491868	1.600316
M	18	6	0	0.249795	1.433813	0.510623
M	25	14	2	0.535234	1.465142	0.125037
F	25	9	1	1.44393	0.158954	1.450151
M	35	14	2	0.536497	1.463171	0.083404
M	50	17	2	0.771203	1.54564	0.354292
F	50	9	1	1.487375	0.367159	1.484995
F	25	14	1	1.512109	0.382936	1.41973
M	56	13	2	0.674961	1.519764	0.459783

Figure 14 : Result of assigning data to clusters


Evaluation_results

Rows 4 Columns 5

Result Description	Average Distance to Other Center	Average Distance to Cluster Center	Number of Points	Maximal Distance to Cluster Center
Evaluation For Cluster No.0	0.611877	0.227945	699	0.551051
Evaluation For Cluster No.1	1.440007	0.343088	537	0.822567
Evaluation For Cluster No.2	0.635199	0.259374	576	0.605027
Combined Evaluation	0.864713	0.272059	1812	0.822567

Figure 15 : K-means clustering model evaluation results

References

[1] Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., ... & Aly, M. (2008, April). Video suggestion and discovery for youtube: taking random walks through the view graph. In Proceedings of the 17th international conference on World Wide Web (pp. 895-904).

[2] Jain, K. N., Kumar, V., Kumar, P., & Choudhury, T. (2018). Movie recommendation system: hybrid information filtering system. In Intelligent Computing and Information and Communication (pp. 677-686). Springer, Singapore.

[3] Keshava, M. C., Reddy, P. N., Srinivasulu, S., & Naik, B. D. Machine Learning Model for Movie Recommendation System.

[4] Wu, C. S. M., Garg, D., & Bhandary, U. (2018, November). Movie recommendation system using collaborative filtering. In 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS) (pp. 11-15). IEEE.

[5] Goyani, M., & Chaurasiya, N. (2020). A Review of Movie Recommendation System. ELCVIA: electronic letters on computer vision and image analysis, 19(3), 18-37.

[6] Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In Journal of physics: conference series (Vol. 1142, No. 1, p. 012012). IOP Publishing.

[7] Mahesh, B. (2020). Machine Learning Algorithms-A Review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386

[8] Bae, S. M., Lee, S. C., & Park, J. H. (2014). Utilization of demographic analysis with imdb user ratings on the recommendation of movies. The Journal of Society for e-Business Studies, 19(3), 125-141.

[9] Kim, M., Jeon, S., Shin, H., Choi, W., Chung, H., & Nah, Y. (2019, June). Movie Recommendation based on User Similarity of Consumption Pattern Change. In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (pp. 317-319). IEEE.

[10] Moler, C., & Little, J. (2020). A history of MATLAB. Proceedings of the ACM on Programming Languages, 4(HOPL), 1-67.

[11] Nawara, D., & Kashef, R. (2021, April). Deploying Different Clustering Techniques on a Collaborative-based Movie Recommender. In 2021 IEEE International Systems Conference (SysCon) (pp. 1-6). IEEE.

[12] Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. Acm transactions on interactive intelligent systems (tiis), 5(4), 1-19.

[13] Zelaya, C. V. G. (2019, April). Towards explaining the effects of data preprocessing on machine learning. In 2019 IEEE 35th international conference on data engineering (ICDE) (pp. 2086-2090). IEEE.