JOURNAL OF SOFT COMPUTING AND DATA MINING VOL.4 NO. 1 (2023) 1-12

JSCDM

http://penerbit.uthm.edu.my/ojs/index.php/jscdm

e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

Adopting Machine Learning in Demographic Filtering for Movie Recommendation System

Lee Jia Yin[1], Noor Zuraidin Mohd Safar[1*], Hazalila Kamaludin[1], Noryusliza Abdullah[1], Mohd Azahari Mohd Yusof[1] , Catur Supriyanto[2]

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

[2]Faculty of Computer Science,
Universitas Dian Nuswantoro, Semarang, INDONESIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2023.04.01.001
Received 04 January 2023; Accepted 02 April 2023; Available online 25 May 2023

Abstract: In this era of big data explosion, humans widely use the movie recommendation system as an information tool. There are two common issues found in the machine learning movie recommendation system that is still undeniable: first, cold start, and second, data sparsity. To minimize the problems, a research study is conducted to find a decision-making algorithm to solve the complex start problem in a movie recommendation system with precise parameters. It involves the implementation of the proposed demographics filtering technique with the k-means clustering method. The research findings present the effects of demographic filtering for movie recommendations. Demographic filtering can group users into clusters based on gender, age group, and occupation. The clusters distribution representative group based on the top 100 results of the experiment. The user with the least distance to the cluster center is chosen as the usual group in that cluster. Three clusters were experimented: Cluster 0, Cluster 1, and Cluster 2. Cluster 0 has a representative group of male, college, or graduate students aged 25 to 34. Cluster 1 has a representative group of females, executive or managerial, aged 25 to 34. Cluster 2 has a representative group of males, sales or marketing aged 35 to 44. It is shown that user from different collection has various preferred movie genre. The preferred movie genre in Cluster 0 is action, adventure, comedy, drama, and war. Cluster 1 has preferred comedy, crime, drama, horror, romance, and sci-fi movie genres. Cluster 2 has chosen action, comedy, drama, film-noir, mystery, and thriller movie genres. This research has contributed to the demographic filtering studies as an alternative solution for future technical development work.

Keywords: Machine learning, recommendation system, demographic filtering, K-means clustering

1. Introduction

The advancement in technology is reaching new heights every day and it is obviously noticeable. To manage a large amount of data, machine learning that builds an analytical model automatically is introduced. People use machine learning to generate a recommender system that predicts most related recommendations using various computational statistics of datasets on the internet [1]. The recommender system is a kind of information filtering system that works to predict 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].

Corresponding author: zuraidin@uthm.edu.my

Interestingly, famous movie streaming service application like Netflix uses CineMatch as their proprietary recommender system since 2000. CineMatch is software that is embedded in the Netflix website where it applies machine learning and data mining to analyze customers' 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 on the certain platform [4].

Generally, there are three main issues with the machine learning implementation in the existing recommendation system. Initially, the cold start problem can be found often in the situation of a new user who 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 which 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 the cold start problem in a movie recommendation system with precise parameters. 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].

2.4 Comparative Study

In this section, a few existing studies are selected and analyzed to have an overview that can be used to relate with the proposed demographic filtering technique in the movie recommendation system.

Table 1 – **The analysis of related studies**

No	Year	Article	Methodology
1	2018	A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System [10]	Hybrid of collaborative filtering and content-based filtering
2	2019	Movie recommendation system using semi-supervised learning [11]	Semi-supervised learning
3	2020	Deep matrix factorization approach for collaborative filtering recommender systems [12]	Deep neural networks
4	2022	Collaborative filtering based on multiple attribute decision making [13]	Multiple attribute decision making (MADM)
5	2022	Demographic Filtering for Movie Recommendation System Using Machine Learning [14]	Demographic-based filtering, decision trees, and random forests

Table 1 shows the analysis of some related studies on the adoption of machine learning in demographic filtering for movie recommendation systems with different findings. The paper "A Hybrid Approach Using Collaborative Filtering and Content-Based Filtering for Recommender Systems" proposed a hybrid approach that combines collaborative filtering and content-based filtering to improve the accuracy of movie recommendations using techniques like clustering, similarity, and classification [10]. The paper "Movie recommendation system using semi-supervised Learning" implemented semi-supervised learning, a machine learning technique that leverages labeled and unlabeled data to improve the accuracy of predictions as a movie recommendation system [11]. While the paper "Deep matrix factorization approach for collaborative filtering recommender systems" proposed the utilization of deep neural networks to learn latent representations of users and items. The proposed deep matrix factorization approach outperforms the other collaborative filtering algorithms in terms of mean absolute error (MAE), root means square error (RMSE), and precision at k [12]. Furthermore, the paper "Collaborative filtering based on multiple attribute decision making" introduced multi-attribute decision making (MADM) to calculate the similarity between users and items using principal component analysis (PCA) based on multiple attributes. This proposed MADM approach outperforms existing collaborative filtering approaches in terms of recommendation accuracy [13]. And last but not least, the paper "Demographic Filtering for Movie Recommendation System Using Machine Learning" used demographic information, machine learning techniques such as decision trees and random forests to predict the rating that a user would give to a movie based on their demographic information and the ratings of similar users. The system performs better than individual models and improve the accuracy and coverage of recommendations [14].

3. Methodology

This chapter studies about the methods and procedures used for the research on machine learning in movie recommendation systems. 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 analysis and experimental setup purposes in this research study are listed 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, 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 [15]. 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 experimental environment for machine learning model training, deployment and, result evaluation for analysis purposes as well.

3.2 Research Framework

The research framework of the demographic filtering movie recommendation system proposed in this paper is shown 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 movie recommendations. K-means clustering is known as a 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 [16]. After the data is scattered into clusters based on 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 has been placed.

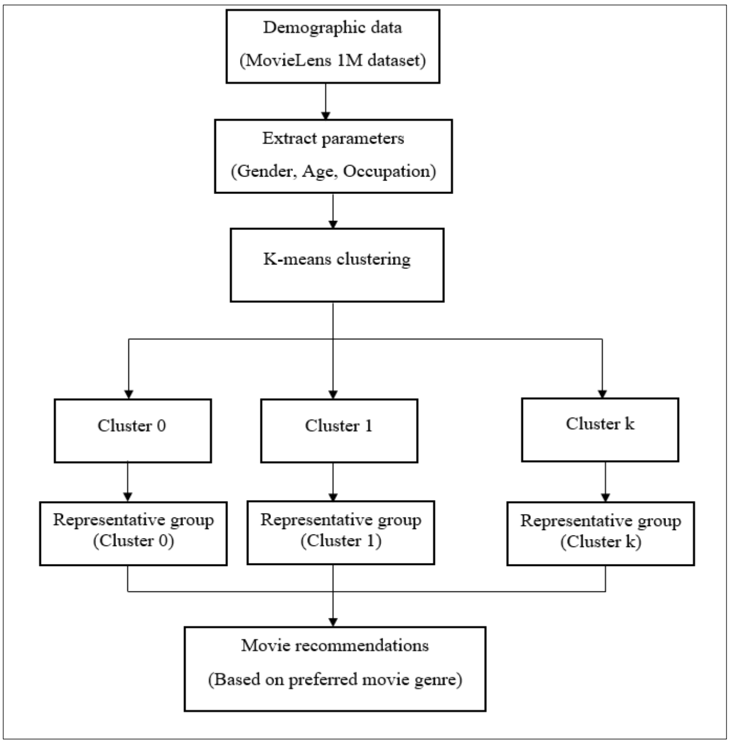


Fig. 1 - Framework of proposed demographic filtering movie recommendation system

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 [17].

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 [18]. In this research study, data pre-processing is done on MATLAB.

In the data pre-processing stage of this research study, function like is missing () is called to find the number of missing values in the dataset imported. Next, remising () is called to remove the missing entries in the dataset imported. Furthermore, removers () 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 Parameters Selection

The parameter is a numerical quantity or attribute used to describe the characteristic of a 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. Table 2 shows the parameters measured and involved. They are the user’s gender, age, and occupation. It also lists all the movie genres used in this study.

Table 2 - Parameters and their attributes

Occupation		Genres	Age	
ID	**Attribute**	**Attribute**	**ID**	**Attribute**
0	Other/Not specified	Action	1	Under 18
1	Academic/Educator	Adventure	18	18 – 24
2	Artist	Animation	25	25 – 34
3	Clerical/Admin	Children’s	35	35 – 44
4	College/Grad student	Comedy	45	45 – 49
5	Customer service	Crime	50	50 – 55
6	Doctor/Health care	Documentary	56	Above 55
7	Executive/Managerial	Drama		
8	Farmer	Fantasy		
9	Homemaker	Film-Noir		
10	K-12 student	Horror		
11	Lawyer	Musical		
12	Programmer	Mystery		
13	Retired	Romance		
14	Sales/Marketing	Sci-Fi		
15	Scientist			
16	Self-employed			
17	Technician/Engineer			
18	Tradesman/Craftsman			
19	Unemployed			
20	Writer			

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 datasets and add components to conduct machine learning experiments designed in the sense of the researcher’s desire using a pipeline.

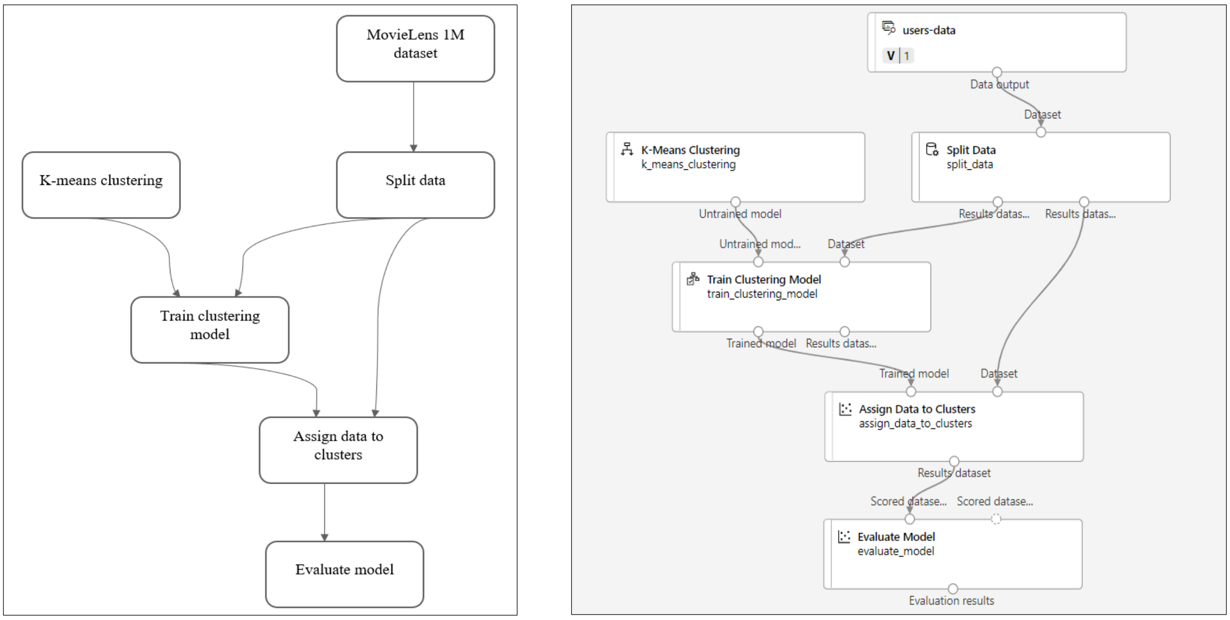


Fig. 2 - Average distance to centroid testing with different number of clusters (a) design; (b) training model

In this study, the pipeline is used and designed following standard machine learning model training procedures as shown in Figure 2. The experiment design aims 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, the 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 the dataset, cleaning missing data, and normalizing data are already completed on MATLAB. The second step in the experiment is to split the data. It is carried out to perform data splitting where 70% of the dataset will be used to train the clustering model and the remaining 30% of the dataset will be assigned to the clusters for evaluation purposes. 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 the training dataset. The third step is to link the 70% dataset into the clustering training model. At the same time, an untrained K-means clustering model is linked to training the clustering model. The next step is to assign the 30% riven data to the trained clustering model. The final step is to evaluate the machine learning clustering model based on the demographic information from MovieLens 1M dataset.

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

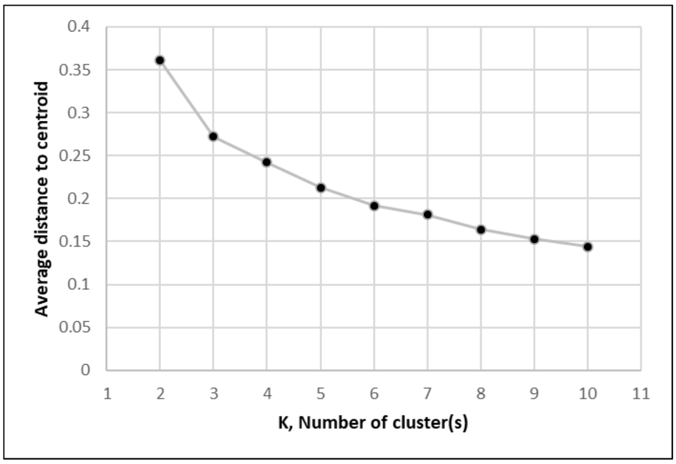


Fig. 3 - Average distance to centroid for different number of clusters

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 have been conducted using tools on MATLAB, Microsoft Excel, and Microsoft Azure Machine Learning Studio. The experiment setup and testing have been 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 data analysis from the outcome of the research. The analysis on the measured parameter and the evaluation on the implementation of the proposed demographic filtering using the K-means clustering are reviewed and discussed.

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 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 users, there are 4331 males and 1709 female contributed to 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 males to female in MovieLens 1M dataset.

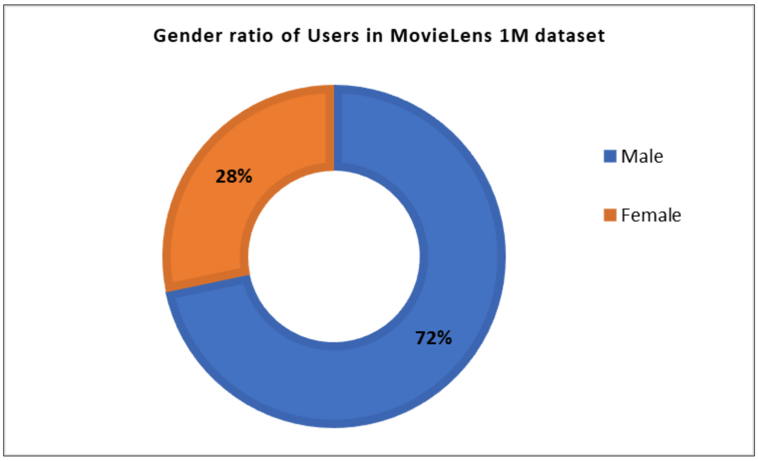


Fig. 4 - Gender ratio of users in MovieLens 1M dataset

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

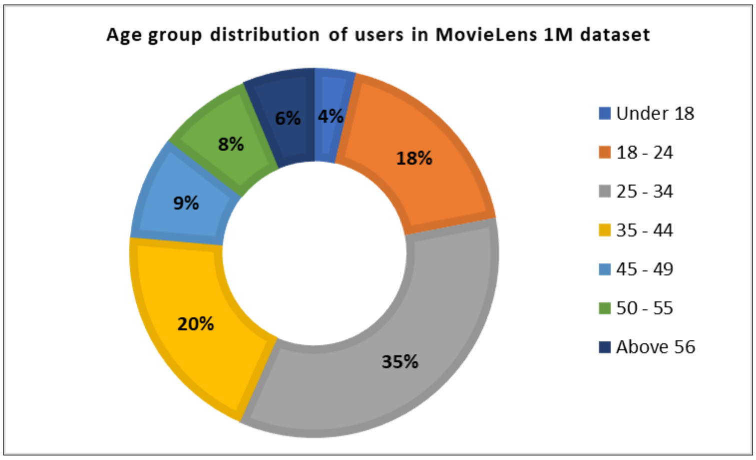
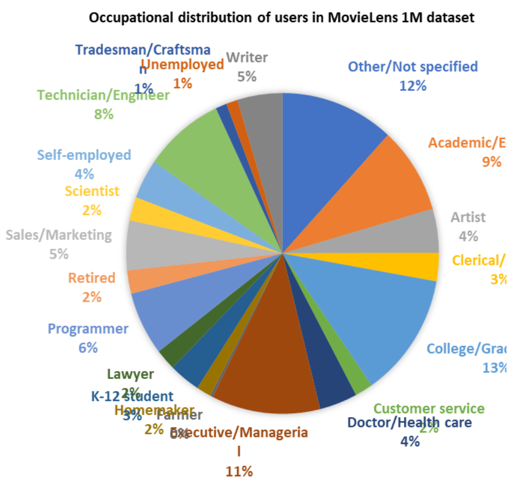


Fig. 5 - Age group distribution of users in MovieLens 1M dataset

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



759	College/Grad student	195	K-12 student
711	Other/Not specified	173	Clerical/Admin
679	Executive/Managerial	144	Scientist
528	Academics/Educator	142	Retired
502	Technician/Engineer	129	Lawyer
388	Programmer	112	Customer service
302	Sales/Marketing	92	Homemaker
281	Writer	72	Unemployed
267	Artist	70	Tradesman/Craftsman
241	Self-employed	17	Farmer
236	Doctor/Health care		

Fig. 6 - Occupational distribution of users in MovieLens 1M dataset

Figure 7 presents the movie rating distribution of the MovieLens 1M dataset. The analysis process on movie ratings is computed by using MATLAB's chart plotting tool. The count for a movie rating of one is 56174, which is the least appointed movie rating. The count for a movie rating of two is 107557. The count for a movie rating of three is 261197. The count for a 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 ratings is 1000209 in MovieLens 1M dataset. The average movie rating from MovieLens 1M dataset is 3.58, where it is left-skewed and negatively skewed.

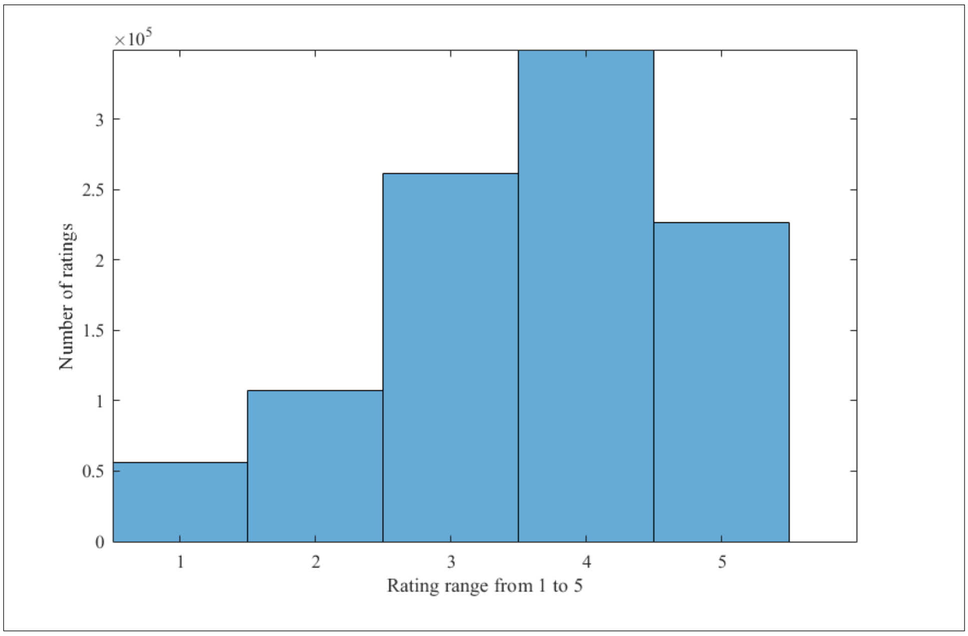


Fig. 7 - The movie rating distribution in MovieLens 1M dataset

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 movies, there are 1161 drama movie, 994 comedy movie, 455 action movie, 217 horror movie, 143 adventure movie, 126 crime movie, 102 documentary movies, 100 thrillers, 83 children's movie, 82 animation movie, 49 romance, movie, 42 sci-fi movies, 35 mystery movie, 33 western movies, 25 musical movies, 25 film-noir movies, 12 war movies, and two fantasy movies. Figure 8 shows the movie genres distribution in MovieLens 1M dataset, where most of the movies are tagged as drama and comedy genre.

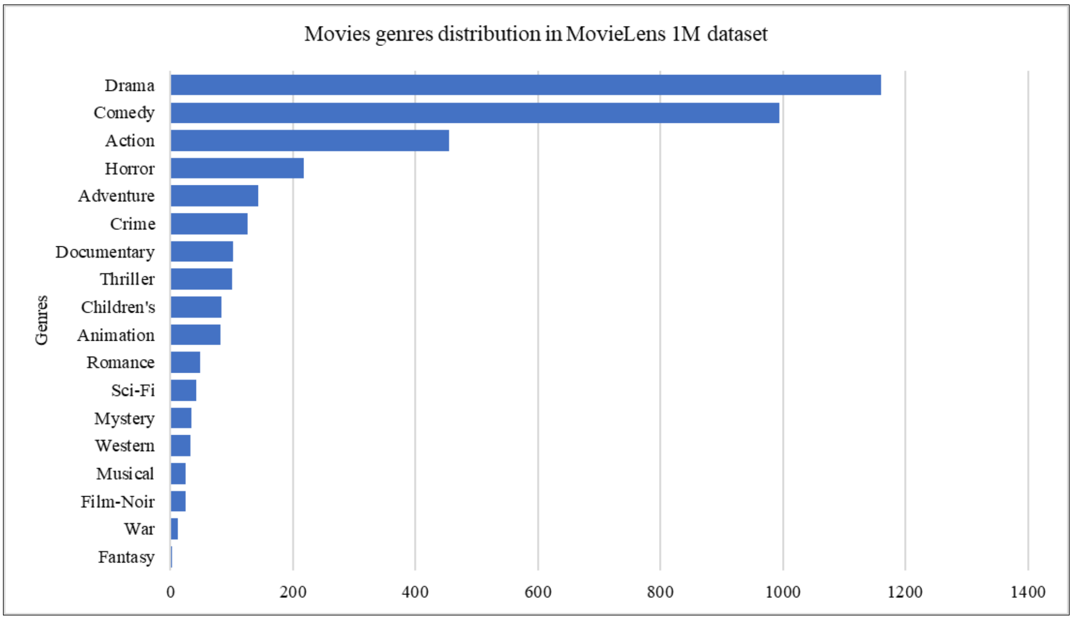


Fig. 8 - The movie genres distribution in MovieLens 1M dataset

4.2 Results Evaluation

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.

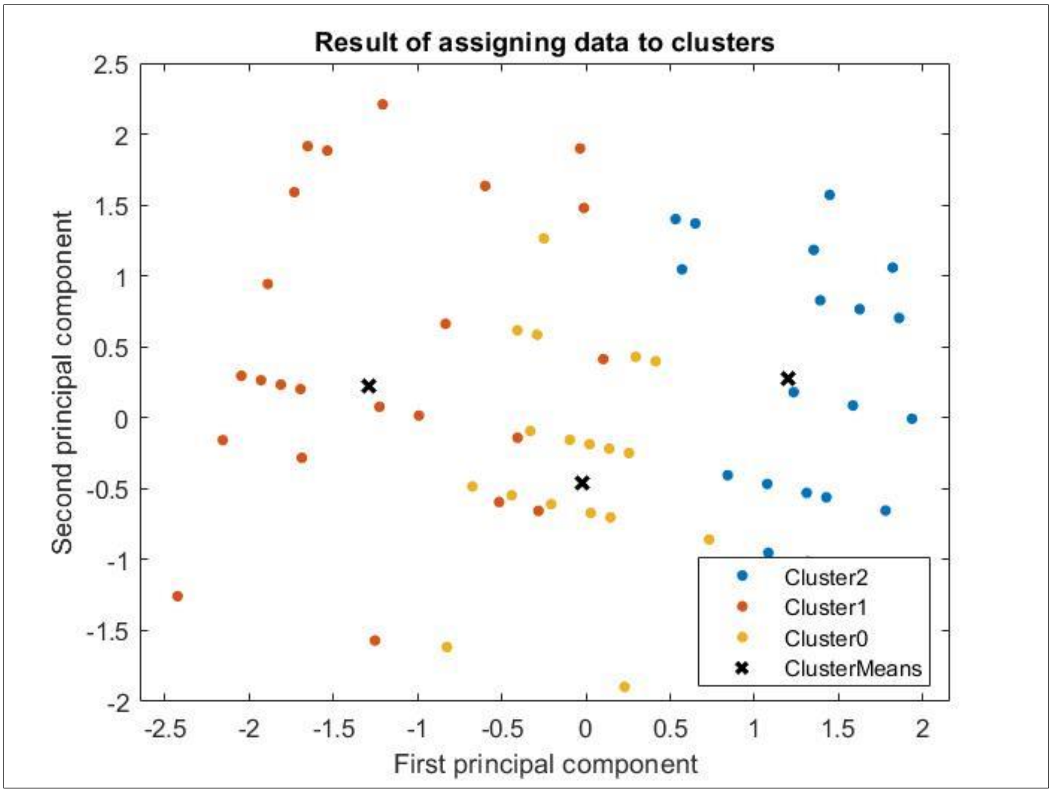


Fig. 9 - Result of assigning data to clusters

Table 3 shows the cluster 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 the cluster center is chosen as the representative group in that cluster. Cluster 0 has a representative group of male, college or grad students aged 25 to 34. Cluster 1 has a representative group of females, executive or managerial, aged 25 to 34. Cluster 2 has a representative group of male sales or marketing aged 35 to 44. Table 4 shows that user from different cluster has various preferred movie genre, whereas the preferred movie genre in Cluster 0 is action, adventure, comedy, drama, and war. Cluster 1 has preferred movie genres of comedy, crime, drama, horror, romance, and sci-fi. And Cluster 2 has preferred movie genres of action, comedy, drama, film-noir, mystery, and thriller.

Table 3 - Clusters distribution

Cluster	Representative group's IDs	Distance to cluster center	Preferred movie genre
0	M, 25, 34	0.091596	Action, Adventure, Comedy, Drama, War
1	F, 25, 34	0.108244	Comedy, Crime, Drama, Horror, Romance, Sci-Fi
2	M, 35, 44	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 4.

Table 4 - 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 5 below, but not limited to them.

Table 5 - 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 of supporting the implementation of a demographic filtering approach in a machine learning movie recommendation system at primary. An effective clustering process based on demographic information can create a productive user categorization or grouping. And based on the clusters, ideally, users in the same cluster will enjoy the recommended movie that comes from a similar genre.

5. Conclusion

The aim of the study is achieved from the outcome of the simple experiment setup to study the adoption of machine learning in movie recommendation systems. The research study establishes demographic filtering with the K-means clustering algorithm to be the solution to the cold start issue on a machine learning movie recommendation system. This paper shall contribute to the development of the recommendation system in the movie industry, considering the performance of demographic filtering in a machine learning movie recommendation system. Nevertheless, the study of machine learning recommendation systems can actually always be implemented in many other applications, but not limited to the movie industry. The objectives of the study were achieved in this project by employing the proposed demographics filtering with the complement of a 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. 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 systems, future works should be considered for further research and development on applying machine learning in movie recommendation systems. Considering demographic filtering is the proposed solution to solve common problems that 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 future works. First, this future work shall generate a strong production that can prove that the proposed demographic filtering is suitable to be implemented in a machine learning movie recommendation system. Second, 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 a parameter to protect user privacy. In this paper, demographic filtering is proposed to be implemented in movie recommendation system when compared to traditional collaborative filtering and content-based filtering. Demographic filtering is projected to solve the cold start problem hence producing accurate movie recommendations for users in a movie recommendation system. After all, with enduring exploitation in the future, there is still a lot of room for improvement in the implementation of artificial intelligence's machine learning in human daily activities.

Acknowledgement

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.

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] Geetha, G.P., Safa, M., Fancy, C., & Saranya, D. (2018). A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System. Journal of Physics: Conference Series, 1000.

[11] Roy, S., Sharma, M., & Singh, S. K. (2019, October). Movie recommendation system using semi-supervised learning. In 2019 Global conference for advancement in technology (GCAT) (pp. 1-5). IEEE.

[12] Lara-Cabrera, R., González-Prieto, Á., & Ortega, F. (2020). Deep matrix factorization approach for collaborative filtering recommender systems. Applied Sciences, 10(14), 4926.

[13] Ya-Jun Leng, Zong-Yu Wu, Qing Lu & Shuping Zhao (2022). Collaborative filtering based on multiple attribute decision making. Journal of Experimental & Theoretical Artificial Intelligence, 34:3, 387-397.

[14] Rastogi, S., Agarwal, D., Jain, J., & Arjun, K. P. (2022, January). Demographic Filtering for Movie Recommendation System Using Machine Learning. In Proceedings of International Conference on Recent Trends in Computing: ICRTC 2021 (pp. 549-557). Singapore: Springer Nature Singapore.

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

[16] 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.

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

[18] 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.