Clustering-Based Recommendation System for Preliminary Disease Detection

Gourav Jain, Indian Institute of Technology, Roorkee, India*
Tripti Mahara, Christ University, India
S. C. Sharma, Indian Institute of Technology, Roorkee, India
Om Prakash Verma, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India
Tarun Sharma, Shobhit University, Gangoh, India

ABSTRACT

The catastrophic outbreak COVID-19 has brought threat to the society and also placed severe stress on the healthcare systems worldwide. Different segments of society are contributing to their best effort to curb the spread of COVID-19. As a part of this contribution, in this research, a clustering-based recommender system is proposed for early detection of COVID-19 based on the symptoms of an individual. For this, the suspected patient’s symptoms are compared with the patient who has already contracted COVID-19 by computing similarity between symptoms. Based on this, the suspected person is classified into either of the three risk categories: high, medium, and low. This is not a confirmed test but only a mechanism to alert the suspected patient. The accuracy of the algorithm is more than 85%.

KEYWORDS

Collaborative Filtering, Coronavirus, COVID-19, Health Recommender Systems, Similarity Measure

INTRODUCTION

SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2), also known as COVID-19 was initially discovered in Wuhan, China, in December 2019(Zoabi et al., 2021). Since then, it has spread around the world and has affected more than 200 countries. According to the figures of WHO, a total of 239,437,517 infected cases and 4879235 deaths are recorded as of October 15th, 2021(WHO Coronavirus (COVID-19) Dashboard). On March 11th, 2020, the World Health Organization (WHO) designated it a “global pandemic” (Coronavirus Disease (COVID-19)). The United States of America (US), the United Kingdom (UK), Brazil, Russia, and India are the top five nations with the most infected COVID-19 cases worldwide. The spread of COVID-19 is unstoppable, and many new forms of it have emerged that are known to be more infectious. Scientists designated one of them as “VUI - 202012/01”(Mishra, 2020).

COVID-19 is a highly contagious disease that is spread from person to person through droplets created by infected people when they sneeze or cough. Since it is a novel virus, scientists
epidemiologists still have a lot to learn about it. Even though vaccines are now available across countries and people are getting vaccinated, they are getting infected by COVID-19 even after vaccination. Hence to curb the spread of this virus, it is important that it is detected at an early stage. Mass testing is one of the most efficient methods for early diagnosis of this virus. It is, however, not a viable solution due to the large population and insufficient medical facilities. Therefore, the use of technology in the medical field can be another way to control the COVID-19 spread. The Recommender System (RS) (G. Jain et al., 2013b) is one such technique of data mining that can be a cost-effective alternative in the early detection of disease. Traditionally an RS collects the user’s historical behaviour, analyses interests, and suggests items according to their interests. Developing a recommender system aims to provide personalized recommendations to customers while selecting an item among a set of products (movies, books, etc.). It became an integral part of the e-commerce business as it has the capability to suggest relevant items to a user depending upon their preferences. Although the idea emerged with the e-commerce domain, it is used in different domains to meet the needs of the users with time (G. Jain et al., 2013a). The health sector is one such domain where RS can be very useful for the primary diagnosis of a disease or to provide users with helpful recommendations based on the symptoms. The system can be based on the patient’s profile or symptoms’ profile (Centers for Disease Control and Prevention). The RS’s CF technique (G. Jain et al., 2020) is the most popular and commonly used neighbourhood-based technique that considers only user-item ratings and ignores the other details. In CF, a similarity measure is used to determine the relationship between users/items, so selecting the proper similarity measure is important in this technique.

In CF, with the ever-increasing volume of data, comparing the user’s preferences with other users/items becomes computationally inefficient and time-consuming. This even needs more deliberation when the data is sparse. Hence, with the objective to make search more efficient in less time, the clustering technique is used. It’s an unsupervised classification method for grouping objects together in such a way that similar objects are clustered together and dissimilar objects are clustered together. In order to determine the similarity among users/items inside the cluster, a similarity measure such as Euclidean distance, cosine, Jaccard, etc. are used. All these methods work fine with numeric data, but in our case, the data is categorical, so these measures are not appropriate. We used Gower’s similarity (All, 2019; Fontecha et al., 2014) to create clusters as it works effectively with any type of data, especially categorical data. It is also used as a metric in clustering algorithms. After creating the cluster using the k-means clustering technique, the similarity between each Suspected Person (SP) is computed with the Corona Patient (CP) by comparing their symptoms. Here, an SP can be defined as a person in which symptoms of Covid-19 are developing and/or he/she may have come in contact with a corona patient but yet has not been confirmed with the COVID-19 disease. The similarities inside a specific cluster are calculated using the improvised Gower (iG) coefficient (Jain et al., 2021), and then based on the calculated similarity, a suspected person is classified into different risk categories: High, Low and Medium. After this, specific recommendations are provided to each risk category of a patient. Although it is not a fool proof way for disease detection and cannot substitute the testing procedure, it is like an alarming system that warns the user about the further course of actions as per the present symptoms. The main contributions of the paper are presented below:

1. Through this paper, we have proposed a recommendation system to predict the risk category of users and provide recommendations.
2. We analysed the efficiency of the improvised Gower’s similarity coefficient with and without clustering approach.

The remainder of this paper is organized as follows. Section 2 presents relevant works on COVID-19 and the Recommendation system. Section 3 explains the proposed algorithm. In Section 4, experiment results are presented. Finally, we conclude the paper in Section 5.
RELATED WORK

The COVID-19 related research has gained momentum in the last year and a half. Some of the methods to predict COVID-19 available in the literature are discussed here.

(Srinivasa Rao & Vazquez, 2020) developed a mobile phone-based web survey to improve the possible case detection of COVID-19 by utilizing machine learning algorithms. (Zoabi et al., 2021) introduced a machine-learning algorithm to predict a positive SARS-CoV-2 infection in an RT-PCR test. To forecast the number of newly infected/recovered patients from COVID-19, (Mojjada et al., 2020) employed some ML models such as the Regression Model (LR), the Lowest Absolute and Selective Shrinking Operator (LASSO), Exponential Smoking (ES), etc. A preliminary test is conducted by (Kamarudin et al., 2021) to forecast the COVID-19 using several techniques such as support vector regression, Gaussian process, linear regression, etc. With the help of several classification techniques such as k-nearest neighbours, decision trees, and random forest, (Muhammad et al., 2020) predicted the likelihood of patients in hospitals in South Korea recovering after contracting COVID-19. Kurniawan et al. (Kurniawan et al., 2020) applied k-means clustering to analyse the correlation between total cases and deaths due to COVID-19. (NAJAR, 2021) presented a hypothetical framework for early detection of COVID-19 from symptomatic information using deep learning models. Automated detection of COVID-19 cases using deep neural networks with X-ray images is presented by (Ozturk et al., 2020).

In the medical sector, the use of data mining techniques is increasing day by day due to precise prediction ability. RS is one such technique that helps the user to provide personalized recommendations based on their interests. In the last few years, a lot of progress has been made on disease prediction using a recommendation system. A brief discussion about a few of them is given here. (Lafta et al., 2016) developed an intelligent recommender system for heart disease patients depending upon the short-term risk prediction. A dynamic recommender system is built by (Nasiri et al., 2016) to predict the susceptible diseases for patients suffering from chronic disease. A smart Health Recommendation System (HRS) employing Restricted Boltzmann Machine (RBM)-Convolutional Neural Network (CNN) was proposed by (Sahoo et al., 2019). This deep learning approach demonstrates how large data analytics can be leveraged to create a useful health recommender engine.

In RS, clustering techniques are used to improve the performance of the system. It helps in finding a group of customers who share the same taste. Several CF-based recommendation algorithms incorporated clustering approaches to minimize the sparsity problem of CF based RS. Some of the techniques are discussed here. (Xiaojun, 2017) Xiaojun proposed an improved clustering-based CF technique to generate users and items clusters using the classical k-means clustering algorithm. A recommendation algorithm based on a combination of user and item clustering is proposed by (Gong, 2010). In this approach, firstly users are clustered based on users’ ratings on items, and then after finding the neighbours of the target users, item clustering is used to produce the recommendations. To solve the k-Means initialization problem, (Tzortzis & Likas, 2014) proposed the MinMax k-means technique. In this, weights are assigned to clusters in proportion to their variance. (Dakhel & Mahdavi, 2011) created a novel collaborative filtering method based on k-means clustering and neighbours voting. Initially in this technique, people are categorized based on their interests using a k-means clustering algorithm and then recommendations are generated for the current user using a novel approach known as a voting algorithm. (Kant et al., 2018) proposed a novel centroid selection approach for the k-means clustering algorithm. This LeaderRank based k-means clustering algorithm overcomes the sparsity issue and improve clustering quality. (Yu & Huang, 2017) created CluCF,
a user and service clusters-based method to overcome the sparsity problem. For the same purpose, (Yusefi Hafshejani et al., 2018) developed a technique by clustering the personality factors.

**METHODOLOGY**

The research focuses on building an RS that predicts the outcome in terms of the severity of contracting the disease. This prediction is based on the symptoms gathered from the suspected person. There are mainly four symptoms usually observed in people who are affected by COVID-19, including fever, cough, difficulty in breathing, and fatigue (Imran et al., 2020). These symptoms might be of varying levels of severity in a patient. For example, some people may experience greater tiredness than others. Similarly, some people may have a high fever, but in others, it may be below. Thus, based on the severity levels, a suspected person (SP) is categorized into Symptomatic, Presymptomatic, Atypical, and Asymptomatic. A person who has all the four symptoms is said to be symptomatic, and a suspected person (SP) who is feeling well today but symptoms may come in a day or two, is considered as presymptomatic. A person who is experiencing just minor symptoms and these symptoms disappear after some time is considered atypical and who does not show any symptoms is classified as asymptomatic (Laila Afifa, 2020). Along with this, a suspected person (SP) is categorized into any of the following three risk categories: High, Medium, and Low. Here, high indicates that the symptoms of the SP falling in the high-risk category are similar to the symptoms of corona patients, while low indicates that the symptoms of the SP falling in the low-risk category are less similar to the symptoms of corona patients. The conceptual model of proposed CF-based RS using the clustering technique is given in Figure 1. The proposed method provides recommendations through the following steps:

1. Creating patient symptoms matrix.
2. Creating clusters using k-means clustering technique.
3. Similarity computation and neighbourhood formation.
4. Generating predictions and providing recommendations.

*Figure 1. Collaborative Filtering based Recommendation System using Clustering technique*
The description of each step is given here in detail.

**Creating Patient Symptoms Matrix**

In this paper, a patient-symptom matrix has been created to store the records of each patient whose sample is displayed in Table 1. In this table, each row represents a patient, each column represents their symptom and the value written inside the box represents the level of intensity of that user’s symptoms.

**Creating Clusters**

Machine Learning techniques (Khanal et al., 2020) are classified into supervised and unsupervised learning techniques. In supervised learning, a model is built to predict the dependent variable from the independent variables using mapping functions, such as classification and regression. Unlike supervised learning, in unsupervised learning, the data is not explicitly labelled into different classes; that means, there are no labels. Out of these two approaches, unsupervised learning is more useful because it has less numerical complexity and it can be used for data analysis in real time. Clustering, anomaly detection, neural networks, etc., are a few examples of unsupervised learning techniques.

In this paper, we use the clustering technique, where objects are clustered in such a way that similar objects are in the same cluster and dissimilar objects are in different clusters. The k-means algorithm attempts to make similar points in the inter-cluster data. Data points are allocated to a cluster depending on the amount of the square distance between the data points and the middle of the cluster (arithmetic mean of all the data points that belong to that cluster). It is preferable to use the lower value of SSE to finalize the clusters. The fewer variations in the clusters, the more homogeneous (similar) data points within the same cluster are formed. This technique traditionally uses Euclidean distance to compute the distance between the cluster’s centroids and the various data points. The Euclidean distance works correctly for the continuous data but it does not give satisfactory results for the categorical data. In this work, the symptoms are represented as non-continuous data, hence the Gower coefficient (G. Jain & Mahara, 2019) is used for the similarity computation between cluster centroid and individual data points. The main purpose of using k-means clustering is to reduce the computational complexity of comparing target users with all other users. By creating the cluster, the search for the patient with similar symptoms is reduced to the cluster that is closest to the target user. The disadvantage of this approach is, the number of clusters should be predefined. For this, we used the most popular Elbow method which determines the number of clusters to be formed after analysing the data. It depicts the cost function value produced by many cluster values. An example of the Elbow method (Masud et al., 2018; Saji, 2021 is shown in Figure 2, where Elbow points suggest the optimum number of clusters.

**Similarity Computation and Neighbourhood Formation**

The effectiveness of any CF-based RS is dependent upon a similarity measure we used. To determine the similarity, this paper uses an improvised Gower’s (iG) similarity coefficient which is the improvised form of the basic Gower’s coefficient. iG eliminates the flaws of basic Gower’s coefficient, which has been discussed in detail in the author’s previous work(G. Jain et al., 2021). The use of this

| Patients | Symptom 1 (Temperature) | Symptom 2 (Cough) | Symptom 3 (Fatigue) | Symptom 4 (Breathing Problem) |
|----------|-------------------------|-------------------|---------------------|-----------------------------|
| Patient - 1 | 5                       | 4                 | 3                   | 1                           |
| Patient – 2 | 1                       | 2                 | 1                   | 0                           |
| Patient – 3 | 2                       | 2                 | 2                   | 1                           |
| Patient - 4 | 5                       | 3                 | 2                   | 0                           |
coefficient is justified because it is capable of working with any type of data. Also, it can handle the dataset easily that contains missing ratings. The similarity using an improvised Gower’s coefficient is calculated as follows:

\[
sim (u_a, u_b)_{\text{improved-Gower's}} = \frac{\sum_{k=1}^{n} D_{u_a u_b}^{\text{improved}} \cdot W_{u_a u_b}}{\sum_{k=1}^{n} W_{u_a u_b}}
\]

(1)

where:

\[
D_{u_a u_b}^{\text{improved}} = 1 - \left| \frac{r_{u_a}^{\text{improved}} - r_{u_b}}{R_{k}} \right|
\]

(2)

In Equation 2:

If \( r_{u_a}^{\text{improved}} - r_{u_b} = R_{k} \)

then \( r_{u_a}^{\text{improved}} - r_{u_b} = r_{u_a} - r_{u_b} \) - random uniform \((0, 0.5)\)

In Eq. 1 and 2, \( r_{u_a} \) and \( r_{u_b} \) is the ratings given by user \( u_a \) and \( u_b \), respectively, and \( R_{k} \) represents the difference between the maximum and minimum rating of each item. After calculating the similarity using similarity measure, the neighbours who have high similarity with the target user within the cluster are selected as the neighbours of that users. In this paper, user means patient, item means symptoms and rating value means the intensity of that symptoms.
Generating Predictions and Providing Recommendations

In the final step, the risk category of each suspected person is predicted, and a set of recommendations is provided to them based on the risk category in which they fall. These categories may be either High, Low and Medium which is depending upon the similarity calculated between the suspected person (target users) and the COVID-19 infected people. Through this paper, the following recommendation is provided o each category of suspected person:

- **High-Risk Category**: This group will constitute the suspected person who has a high similarity with the corona-affected patient. Patients falling in this category are advised to visit the nearest testing centre for further tests immediately.

- **Medium-Risk Category**: The individuals of this group have neither very high nor very low similarity with the corona-affected patients. So, they are advised to check their symptoms regularly and visit the testing centre if the symptoms persist for three days. Also, maintain social distancing at home till the symptoms start to subside.

- **Low-risk Category**: Individuals in this category have mild to minor symptoms. But, if they feel any changes in the symptoms, they are advised to recheck their symptoms and visit the doctor kindly.

The algorithm for the proposed measure is given in Algorithm 1.

**EXPERIMENTAL SETUP**

In our previous work (G. Jain et al., 2021), we present the conceptual model as no dataset is publicly available to date. This research first applies the iG coefficient without clustering on the publicly available dataset and then compares its performance iG with the clustering approach. An improvised Gower’s coefficient is used to calculate the similarity among patients.

**Dataset**

The dataset contains various fields like age, gender, travel history, loss of sense of smell, etc., but we utilized the only four main symptoms: temperature, cough, fatigue (Tiredness), and breathing problems of each suspected person. In the dataset, users are classified into three categories: high, low, and medium, based on their symptoms. Here, zero represents the low-risk category of users, 1 represents the medium category of users, and 2 is used to represent the high-risk categories of users. The dataset is publicly available at (I. Jain, 2020) and has a small number of users 127. Out of which, some are used for training purposes and the rest are used for testing purposes. We tested our approach at different-2 training testing ratios.

**Evaluation Criteria**

The Confusion Matrix, Precision, Recall, and F1-score (Jayaswal, 2020) are used to evaluate the performance of the proposed system. The description of each of them is given below.

Algorithm 1. Algorithm for the proposed measure iG with clustering approach

1. Create patient symptom rating matrix
2. Create clusters using k-means clustering algorithm
3. Split each clustered part into a Training - Testing set
4. For each training testing set **repeat steps 5 to 7**
5. Calculate the similarity between suspected persons and the corona patients using improvised Gower’s coefficient
6. Find the neighbours(k) of each suspected person
7. Generate the prediction for each suspected person
8. end for
9. Evaluate the performance by performance metrics
Confusion Matrix

The confusion matrix is used to calculate the precision, recall and F1-score. As shown in Table 2, it is a 2x2 binary classification matrix with real values on one axis and predicted values on the other. True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) are the terminology used in the confusion matrix. Here, TP means, when the predicted and actual numbers both are positive, this is referred to as True Positive. True Negative means when both predicted and actual values are negative, it is TN. False Positive occurs when the model wrongly predicts the negative class. It signifies that a model predicted a positive class, but the actual class turned out to be negative. False Negative (FP) is the total opposite of this. It arises when a model wrongly predicts the positive class.

Precision

Precision is a measure that indicates how many of the positive predictions made are correct. To calculate this, True Positive (TP) and False Positive (FP) values are required:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

Recall

The recall is a measure that shows how many of the positive cases the model correctly predicted, overall, the positive cases in the data. A recall is measured using True positive (TP) and False Negative (FN) values. In the medical application, we don’t want to miss any patient, therefore, we focus on having a high recall:

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

F1-Score

The F1 score sometimes known as F-score is a measure of testing accuracy. It is the harmonic mean of precision and recall. The F1 score considers both False Positives and False Negatives, as a result, it works well with an unbalanced dataset:

\[
\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

RESULTS AND DISCUSSION

To show the effectiveness of the proposed method “iG with clustering approach”, we first evaluate the performance of “iG without clustering approach” on the above-discussed dataset. The experimental

Table 2. Confusion Matrix

| Predicted Values | Actual Values |
|------------------|---------------|
|                  | Positive      | Negative     |
| Positive         | TP            | FP           |
| Negative         | FN            | TN           |
results are evaluated in terms of Precision, Recall, and F-score. The values of TP, TN, FP, and FN are shown using a confusion matrix for each training testing ratio in Table 3. In order to improve the prediction accuracy, we apply clustering technique with iG on the same dataset. The experimental results of the proposed method “iG with clustering approach” for different training testing ratios are displayed using Table 4.

Table 3. Confusion Matrix for iG without clustering approach for various Training-Testing ratios

| Training (60)-Testing (40) | Actual Values |   |   |
|----------------------------|---------------|---|---|
| Predicted Values           |               |   |   |
| Positive                   | 25            |   |   |
| Negative                   | 7             |   |   |
| Training (70)-Testing (30) | Actual Values |   |   |
| Predicted Values           |               |   |   |
| Positive                   | 18            |   |   |
| Negative                   | 6             |   |   |
| Training (80)-Testing (20) | Actual Values |   |   |
| Predicted Values           |               |   |   |
| Positive                   | 10            |   |   |
| Negative                   | 4             |   |   |
| Training (90)-Testing (10) | Actual Values |   |   |
| Predicted Values           |               |   |   |
| Positive                   | 5             |   |   |
| Negative                   | 3             |   |   |

Table 4. Confusion Matrix for iG with clustering approach for various Training-Testing ratios

| Training (60)-Testing (40) | Actual Values |   |   |
|----------------------------|---------------|---|---|
| Predicted Values           |               |   |   |
| Positive                   | 27            |   |   |
| Negative                   | 5             |   |   |
| Training (70)-Testing (30) | Actual Values |   |   |
| Predicted Values           |               |   |   |
| Positive                   | 18            |   |   |
| Negative                   | 4             |   |   |
| Training (80)-Testing (20) | Actual Values |   |   |
| Predicted Values           |               |   |   |
| Positive                   | 11            |   |   |
| Negative                   | 3             |   |   |
| Training (90)-Testing (10) | Actual Values |   |   |
| Predicted Values           |               |   |   |
| Positive                   | 6             |   |   |
| Negative                   | 2             |   |   |
For better performance, the precision value should be low, while the recall and F-value should be high. From the experimental results shown in Figure 3-5, it is obvious that our proposed method iG with clustering achieved this goal successfully. The precision values are lower for every training testing ratio of the dataset as compared to when iG without clustering is applied. The proposed technique also has higher recall and F-score values than iG without clustering approach. The experimental results indicate that improvised Gower’s coefficient with clustering predicts the risk more correctly than iG without clustering approach.

CONCLUSION AND FUTURE SCOPE OF WORK

SARS started a never-ending series of pandemic diseases in 2003, followed by Influenza (H1N1)-2009, MERS-2012, Ebola-2014, Zika Virus - 2016, and finally, COVID-2019 came. This chain has made
the entire planet fearful of any incoming virus. Viruses like coronavirus leave significant marks on the behaviour of people in many aspects including, psychological and social behaviours. The world after corona needs something to keep them feeling safe from any incoming epidemic diseases or the same epidemic with different shapes. Thus, society needs a system that predicts the disease at an early stage. Therefore, through this paper, we have proposed a recommendation system to predict the risk category of users and provide recommendations. To calculate the similarity, among patients an improvised Gower coefficient $iG$ is used inside the cluster that is created using the k-means clustering technique. The experimental results indicate that $iG$ with clustering gives better results than $iG$ without clustering approach.

One of the major limitations of this work is the unavailability of a suitable dataset. Future work includes testing our work on a larger dataset as well as developing an application so that a person can use it and make decisions accordingly. In addition, this work can be extended further by using optimization techniques in conjunction with the proposed work to make the prediction more accurate.
REFERENCES

All, U. T. C. (2019). Extending Gower’s General Coefficient of Similarity to Ordinal Characters. https://www.jstor.org/stable/1224438

Angappan, K. (2021). An Efficient way of Predicting Covid-19 using Machine and Deep Learning Algorithms. 10.4108/eai.7-6-2021.2308785

Arora, R. K., Gupta, M. K., & Bhati, B. S. (2021). Analysis of Various COVID-19 Prediction Techniques. IEIE Transactions on Smart Processing & Computing, 10(4), 323–329. doi:10.5573/IEIEESPC.2021.10.4.323

Coronavirus disease (COVID-19). (n.d.). Retrieved February 24, 2021, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Dakhel, G. M., & Mahdavi, M. (2011). A new collaborative filtering algorithm using K-means clustering and neighbors’ voting. Proceedings of the 2011 11th International Conference on Hybrid Intelligent Systems, HIS 2011, 179–184. doi:10.1109/HIS.2011.6122101

Fontecha, J., Hervás, R., & Bravo, J. (2014). Mobile services infrastructure for frailty diagnosis support based on Gower’s similarity coefficient and treemaps. Mobile Information Systems, 10(1), 127–146. doi:10.1155/2014/728315

Gong, S. (2010). A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering. 10.4304/jsw.5.7.745-752

Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, M. S., Ali, K., John, C. N., Hussain, M. I., & Nabeel, M. (2020). AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Informatics in Medicine Unlocked, 20, 100378. doi:10.1016/j.imu.2020.100378 PMID:32839734

India: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data. (n.d.). Retrieved January 18, 2022, from https://covid19.who.int/region/searo/country/in

Jain, G., & Mahara, T. (2019). An efficient similarity measure to alleviate the cold-start problem. In 2019 Fifteenth International Conference on Information Processing (ICINPRO). doi:10.1109/ICINPro47689.2019.9092250

Jain, G., Mahara, T., & Sharma, S. C. (2021). A Collaborative Filtering-Based Recommendation System for Preliminary Detection of COVID-19. doi:10.1007/978-981-16-1696-9_3

Jain, G., Mahara, T., & Tripathi, K. N. (2020). A Survey of Similarity Measures for Collaborative Filtering-Based Recommender System. Advances in Intelligent Systems and Computing, 1053, 343–352. doi:10.1007/978-981-15-0751-9_32

Jain, G., Mishra, N., & Sharma, S. (2013a). A Survey on Recommendation Techniques in Numerous Domains. International Journal of Computers and Applications, 67(25), 26–30. doi:10.5120/11745-7379

Jain, G., Mishra, N., & Sharma, S. (2013b). CRLRM: Category based Recommendation using Linear Regression Model. Proceedings - 2013 3rd International Conference on Advances in Computing and Communications, ICACC 2013, 17–20. doi:10.1109/ICACC.2013.11

Jain, I. (n.d.). Covid19 Patient Symptoms. 2020. Retrieved May 18, 2021, from https://www.kaggle.com/bitsofishan/covid19-patient-symptoms

Jayaswal, V. (n.d.). Performance Metrics 2020. Retrieved June 17, 2021, from https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262

Kamarudin, A. N. A., Zainol, Z., Kassim, N. F. A., & Sharif, R. (2021). Prediction of COVID-19 cases in Malaysia by using machine learning: A preliminary testing. 2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021. doi:10.1109/WiDSTaif52235.2021.9430222

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

Khanal, S. S., Prasad, P. W. C., Alsadoon, A., & Maag, A. (2020). A systematic review: Machine learning based recommendation systems for e-learning. Education and Information Technologies, 25(4), 2635–2664. doi:10.1007/s10639-019-10063-9
Kurniawan, R., Abdullah, S. N. H. S., Lestari, F., Nazri, M. Z. A., Mujahidin, A., & Adnan, N. (2020). Clustering and Correlation Methods for Predicting Coronavirus COVID-19 Risk Analysis in Pandemic Countries. 2020 8th International Conference on Cyber and IT Service Management, CITSM 2020. doi:10.1109/CITSM50537.2020.9268920

Lafta, R., Zhang, J., Tao, X., Li, Y., & Tseng, V. S. (2016). An intelligent recommender system based on short-term risk prediction for heart disease patients. Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, 102–105. doi:10.1109/WI-IAT.2015.47

Laila Afifa. (2020). Here Are 3 Classifications of Asymptomatic COVID-19 Patients - Life en.tempo. https://en.tempo.co/read/1339559/here-are-3-classifications-of-asymptomatic-covid-19-patients

Masud, M. A., Huang, J. Z., Wei, C., Wang, J., Khan, I., & Zhong, M. (2018). I-nice: A new approach for identifying the number of clusters and initial cluster centres. Information Sciences, 466, 129–151. doi:10.1016/j.ins.2018.07.034

Mishra, M. (2020). New COVID-19 Strain Reaches India: Does It Spread Easily? Vaccines Won’t Work? Here Are All Your Questions Answered. https://www.moneycontrol.com/news/trends/health-trends/new-covid-19-strain-reaches-india-does-it-spread-easily-vaccines-wont-work-here-are-all-your-questions-answered-6280581.html

Mojjada, R. K., Yadav, A., Prabhu, A. V., & Natarajan, Y. (2020). Machine learning models for covid-19 future forecasting. Materials Today: Proceedings. Advance online publication. doi:10.1016/j.matpr.2020.10.962 PMID:33318952

Muhammad, L. J., Islam, Md. M., Usman, S. S., & Ayon, S. I. (2020). Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients’ Recovery. SN Computer Science, 1(4). 10.1007/s42979-020-00216-w

Nasiri, M., Minaei, B., & Kiani, A. (2016). Dynamic Recommendation: Disease Prediction and Prevention Using Recommender System. International Journal of Basic Science in Medicine, 1(1), 13–17. doi:10.15171/ijbsm.2016.04

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Rajendra Acharya, U. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in Biology and Medicine, 121, 103792. Advance online publication. doi:10.1016/j.compbiomed.2020.103792 PMID:32568675

Sahoo, A. K., Pradhan, C., Barik, R. K., & Dubey, H. (2019). DeepReco: Deep Learning Based Health Recommender System Using Collaborative Filtering. Computation, 7(2), 25. 10.3390/computation7020025

Saji, B. (n.d.). Elbow Method 2021. Retrieved May 12, 2021, from https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

Srinivasa Rao, A. S. R., & Vazquez, J. A. (2020). Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. Infection Control and Hospital Epidemiology, 41(7), 826–830. doi:10.1017/ice.2020.61 PMID:32122430

Centers for Disease Control and Prevention. (n.d.). Symptoms of COVID-19. Retrieved April 18, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

Tzortzis, G., & Likas, A. (2014). The MinMax k-Means clustering algorithm. Pattern Recognition, 7(47), 2505–2516. doi:10.1016/j.patcog.2014.01.015

Wang, D., Mo, J., Zhou, G., Xu, L., & Liu, Y. (2020). An efficient mixture of deep and machine learning models for COVID-19 diagnosis in chest X-ray images. PLoS One, 15(11), e0242535. doi:10.1371/journal.pone.0242535 PMID:33201919

Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos, Solitons, and Fractals, 139, 110058. doi:10.1016/j.chaos.2020.110058 PMID:32834611

Xiaojun, L. (2017). An improved clustering-based collaborative filtering recommendation algorithm. Cluster Computing, 20(2), 1281–1288. doi:10.1007/s10586-017-0807-6

Yu, C., & Huang, L. (2017). CluCF: A clustering CF algorithm to address data sparsity problem. Service Oriented Computing and Applications, 11(1), 33–45. doi:10.1007/s11761-016-0191-8
Yusefi Hafshejani, Z., Kaedi, M., & Fatemi, A. (2018). Improving sparsity and new user problems in collaborative filtering by clustering the personality factors. *Electronic Commerce Research, 18*(4), 813–836. doi:10.1007/s10660-018-9287-x

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digital Medicine, 4*(1), 1–5. 10.1038/s41746-020-00372-6