Musical Instruments Recommendation System Using Collaborative Filtering and KNN

Alfriska Deviane Puspita[1], Vynska Amalia Permadi [2], Aliza Hanum Anggani [3], Edwina Ayu Christy [4]
[1234] Department Informatics, Faculty of Industrial Engineering, UPN "Veteran", Yogyakarta, Indonesia, 55283
Email: 123170108@student.upnyk.ac.id[1]; vynspermadi@upnyk.ac.id[2], 123170090@student.upnyk.ac.id[3], 123170095@student.upnyk.ac.id[4]

ABSTRACT

Introduction – The trend of e-commerce and online shopping has offered customers more product choices, but it also resulted in information overload. Nowadays, users are equipped with technology that allows websites to automatically deliver products that they may be interested in so that they can easily locate their favorite items from enormous options. To automate the recommendation process, recommender systems are created and built. This research creates a musical instrument recommendation system based on user reviews.

Methodology/Approach – In this paper, we design and implement a recommendation system that combines the k-Nearest Neighbor (kNN) algorithm with a collaborative filtering framework. Collaborative filtering is chosen in this case because of its capability of providing new information to users by collecting information that has been obtained from the other users. Furthermore, kNN is considered as a suitable combination in this case since this method is relatively simple and able to find the similarity of objects being compared.

Findings – To evaluate this study, the recommendation results are evaluated using the Root Mean Square Error (RMSE) calculation method, and the RMSE result obtained is 0.8734 for schema that divides dataset into 70% data train and 30% dataset using KNNWith Means with pearson measurements, and the MAE (Mean Absolute Error) result obtained is 0.5998 with schema 60% data train and 40% data test using KNNBasic algorithm with cosine similarity.

Originality/ Value/ Implication – We present experimental results of consolidating the kNN algorithm in the collaborative filtering framework using Amazon's musical instrument dataset. Furthermore, we can see that kNN together with a collaborative filtering algorithm performs a satisfactory outcome.

Keywords: Collaborative Filtering, KNN, Recommendation System

INTRODUCTION

The emergence and spread of the internet have provided much information to users. The amount of information provided makes it difficult for users to find information that suits their needs. This information overload problem can be solved using a recommendation system that will provide information and products recommendations according to users' needs. Furthermore, vast amounts of information sometimes make users not understand the product they are looking for, so they only use the keyword according to their understanding. This condition triggers a new problem, namely a mismatch between the keywords entered by the user and the primary data owned by the system, so the result becomes less accurate.

This research will create a musical instrument recommendation system to make it easier for users to find musical instruments that suit their needs. Several methods can be used for recommendation systems, including Content-Based Filtering, Collaborative Filtering, and Hybrid filtering methods. Although several methods can be used, the collaborative filtering method is more companies often use them because it involves more user assessment.

Collaborative Filtering is considered capable of producing better predictions than the content-based filtering method because the collaborative filtering method analyzes the user's search history and compares with other users, and then makes recommendations in order (Konstan et al., 1997). Meanwhile, content-based Filtering takes user information to find similar products and recommends them in order. However, this content-based filtering method has a shortcoming in accuracy and precision.

Several previous studies align with this research, one of which is the music recommendation systems based on K-NN conducted by (Li & Zhang, 2018). This research proves that using K-NN, smaller error values obtained and resulted in the RMSE values of 0.870 and MAE of 0.683. The second research is about film recommendations using Collaborative Filtering (Gupta et al., 2020). This research got TP rate 0.761, precision 0.782, and F1 0.772. The third research is about movie recommendation systems using K-Means clustering and K-Nearest Neighbours conducted by (Ahuja et al., 2019).

Based on several previous studies, the method used in this recommendation system is collaborative Filtering and KNN. KNN algorithm and collaborative Filtering are used to improve the performance of recommendation systems. In addition, the variable used to obtain recommendations in this study is a rating (overall), so the KNN algorithm is appropriate to use in the collaborative filtering method.

LITERATURE REVIEW

The previous research that became the reference for this research was research written by Gang Li and Jingjing Zhang. The title of this research is "Music personalized recommendation system based on improved KNN algorithm". In this study, the authors get the movielens data by crawling from one platform on the internet. This study of KNN-Improved proved to be an algorithm with a lower error rate than other algorithms. This research uses algorithm KNN based on collaborative Filtering and combined with the Baseline algorithm. Root Mean Square Error (RMSE) value from the KNN-Improved algorithm was 0.870, and Mean Absolute Error (MAE) was 0.683. When using a large dataset, KNN-Improved also has a shorter running time than KNN.

The title of the second research is "Movie Recommender System Using Collaborative Filtering" from Meenu Gupta, Aditya Thakkar, and Aashish. Evaluation of this study is seen from the value of TP rate, precision, and F1. Content-based got TP rate 0.591, precision 0.501, and F1 0.528. Collaborative filtering got TP rate 0.761, precision 0.782, and F1 0.772. From this second study, we can

conclude that collaborative filtering methods show better results than content-based Filtering.

Alisha Baskota and Yiu-Kai Ng wrote the third study. The title of this study is "A Graduate School Recommendation System Using the Multi-Class Support Vector Machine and KNN Approaches". The precision value in this study was 0.54, recall 0.47, F-measure 0.50, and accuracy 0.55.

- Collaborative Filtering

Collaborative Filtering is a technique for recommending items to users, based on other user's reviews. *Collaborative filtering* will look for users who have similar or even the same preferences as the targeted user. In this kind, if a person A's traits are just like a few different person B then, the goods that B preferred are encouraged to A. As a statement, we will say, "the customers who like merchandise just like you furthermore mght preferred the ones merchandise". So right here we propose the use of the similarities among customers.

The motivation for collaborative filtering comes from the concept that human beings regularly get the first-rate hints from a person with tastes just like themselves. Collaborative filtering encompasses strategies for matching human beings with comparable pastimes and making hints in this basis. Collaborative filtering algorithms regularly require (1) users' lively participation, (2) an clean manner to symbolize users' pastimes, and (3) algorithms which are capable of healthy human beings with comparable pastimes.

The following outlines is the mechanism of the *collaborative filtering* recommender system :

(1) Create a database containing various kinds of products that get high ratings from customers.
(2) The results of new transactions from customers will be searched for the most similar to the previous data that has been saved to the database.
(3) Users will get recommended results from previous similarity searches.(Erlangga & Sutrisno, 2020)

- K-Nearest Neighbor algorithm

The K-Nearest Neighbour algorithm is still prevalent in recommendation systems. This algorithm is one of the most common algorithms for collaborative Filtering (Muneer et al., 2021). Below is the basic KNN for defined as :

$$\widehat{r_{ui}} = \frac{\sum_{j \in N_u^k(i)} sim(u,v).r_{uj}}{\sum_{j \in N_u^k(i)} sim(u,v)} \quad \text{..... (1)}$$

Or

$$\widehat{r_{ui}} = \frac{\sum_{v \in N_i^k(u)} sim(u,v).r_{vi}}{\sum_{v \in N_i^k(u)} sim(u,v)} \quad \text{.....(2)}$$

The k-Nearest Neighbor algorithm with means model uses similarities between users and or items. This similarity becomes the weight for predicting the recommended rating given. This KNN taking into account mean ratings of user

$$\widehat{r_{ui}} = \mu_i + \frac{\sum_{j \in N_u^k(i)} sim(i,j).(r_{uj} - \mu_j)}{\sum_{j \in N_u^k(i)} sim(i,j)} \quad \text{..... (3)}$$

Or

$$\widehat{r_{ui}} = \mu_u + \frac{\sum_{v \in N_i^k(u)} sim(u,v).(r_{vi} - \mu_v)}{\sum_{v \in N_i^k(u)} sim(u,v)} \quad \text{..... (4)}$$

In general, to calculate the similarity in the K-Nearest Neighbors algorithm, the Euclidean distance is used as the distance calculation method. In this study, there are 4 similarity measurements, where the distance is calculated in a euclidean vector space(Arai et al., 2020).

(1) Cosine Similarity

$$S_{uv} = cos(u,v) = \frac{u.v}{||u||2||v||2} = \frac{\sum_{i \in I_{uv}} r_{ui}.r_{vi}}{\sqrt{\sum_{i \in I_{uv}} r_{ui}^2}\sqrt{\sum_{i \in I_{uv}} r_{vi}^2}} \quad \text{...(5)}$$

(2) Mean Squared Difference (MSD)

$$S_{uv} = \frac{1}{MSD\ (u,v) + 1'} \quad \text{……..(6)}$$

with

$$MSD(u,v) = \frac{1}{|I_{uv}|}.\sum_{i \epsilon I_{uv}} (r_{ui} - r_{vi})^2 \quad \text{….(7)}$$

(3) Pearson Correlation Coefficient

$$S_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - \mu_u)(r_{vi} - \mu_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \mu_u)^2}.\sqrt{\sum_{i \in I_{uv}} (r_{vi} - \mu_v)^2}} \quad (8)$$

Or

$$S_{ij} = \frac{\sum_{u \in U_{ij}} (r_{ui} - \mu_i)(r_{uj} - \mu_j)}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - \mu_i)^2}.\sqrt{\sum_{u \in U_{ij}} (r_{uj} - \mu_v)^2}} \quad (9)$$

(4) Pearson Correlation Coefficient – Baseline

$$S_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - b_u)(r_{vi} - b_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - b_{ui})^2}.\sqrt{\sum_{i \in I_{uv}} (r_{vi} - b_{vi})^2}} \quad (10)$$

Or

$$S_{ij} = \frac{\sum_{u \in U_{ij}} (r_{ui} - b_i)(r_{uj} - b_j)}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - b_{ui})^2}.\sqrt{\sum_{u \in U_{ij}} (r_{uj} - b_{uj})^2}} \quad (11)$$

- Evaluation Metrics

The main objective of this recommendation system experiment is to evaluate the efficiency and pertormance of the proposed recommendation system which is to musical instrument recommendations. There are two metrics that can be measured to evalute the recommendation system, named *Mean Absolute Error (MAE)* and *Root Mean Absolute Error (RMSE)* (Muneer et al., 2021). The smaller the RMSE and MAE values, the more efficient the recommendation system performance.

Mean Absolute Error (MAE)

MAE is defined as the difference between the actual value and the predicted value. MAE can be calculated by the formula below (Muneer et al., 2021)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \quad \text{.....(12)}$$

Root Mean Square Error (RMSE)

The RMSE result close to 0 is an indication of the high accuracy of a model. RMSE can be calculated by the formula below (Muneer et al., 2021):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(p_i - r_i)^2} \quad \text{.....(13)}$$

METHOD

This study focuses on getting recommendations in the form of text by adapting the Cross-Industry Standard Process for Data Mining (CRISP-DM) research methodology. CRISP-DM is served as a standard nonproprietary methodology for data mining, consists of 6 main stages, as shown in Figure 1.

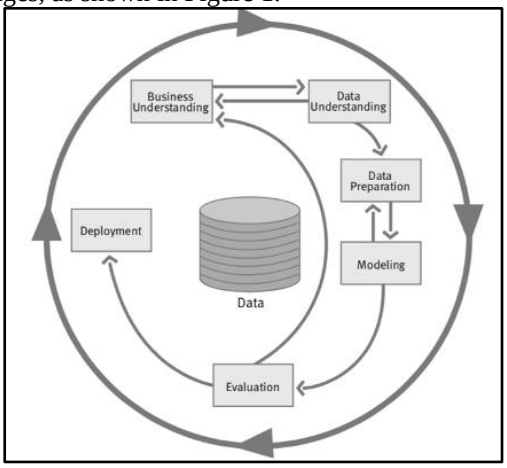

Figure 1. Phases of the CRISP-DM

BUSINESS UNDERSTANDING

The purpose of this study is to identify the similarity of preferences between one user and another. In addition, a question arises that must be resolved and understood through this study, namely "How to create a collaborative filtering recommendation system for musical instruments selection".

The business understanding process can be done repeatedly according to the flow in the picture above. This stage is intended to review the suitability of the data, as well as the model made. keep in mind, that the business understanding process is a process that continues to remind the purpose of making the recommendation system.

DATA UNDERSTANDING

At this stage, the researcher downloaded data from the *kaggle.com/eswarchandt/amazon-music-reviews* page. The downloaded musical instruments data consists of 10262 musical instruments data, with several features: *reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime,* and *reviewTime.* The *reviewerID* column represents the id owned by each consumer who has given a review of the product. The a*sin* column contains information about the product id of each musical instrument. *ReviewerName* is the name of people who have given reviews for the product. The *reviewText* column describes consumer opinion about the product. *Overall* is the rating of the musical instruments. The *reviewTime* column shows the time when consumers write reviews about the product.

The data used are explored under research needs. The researcher checked the data size, checked the null value, the lowest overall highest, the number of each feature, did the data visualization and checked the data type. This stage can consider the researchers about the features that are selected as datasets.

.

DATA PREPARATION

The data that has been understood is then prepared to prepare the dataset. A dataset is a collection of data, or from the past that is processed into meaningful information. This stage begins by creating a new data frame consisting of the initial data grouped by *asin*. Next, the data used is only *asin*, which has an lambda value greater than 20. The remaining data is only 3577 data. To make the modelling, the columns *Overall* that are processed are only *reviewerID*, *asin*, and *Overall*. Figure 2 shows the contents of the top data frame. The lowest overall value in this data frame is 3.8 with 20 data, and the highest is 5.0 with 163 data. This data frame is used as a dataset for modelling in this study.

This stage started with making a new data frame consisting of the original data group by *asin* and filtered with min lambda 20. At this point, the data frame consists of 3577 data.

	reviewerID	asin	overall
41	AA5TINW2RJ195	B000068NW5	5.0
42	ABC68JUCPTVOE	B000068NW5	5.0
43	A3W2E6S24BTXXK	B000068NW5	5.0
44	A3872Y2XH0YDX1	B000068NW5	5.0
45	A398X9POBHK69N	B000068NW5	4.0

Figure 2. Data

MODEL BUILDING

At this stage, the researchers coded and modelled the musical instruments recommendation system. Researchers use Python 3.9 for coding and take advantage of several existing libraries. To get the recommendation for the closest n-item similarity, the researcher uses the *K-Nearest Neighbor algorithm*. The k - Nearest Neighbor algorithm aims to classify objects into a predetermined class based on the object's distance to several K objects previously classified. This research uses KNN to better result at the evaluation because it uses the normalized data before. This algorithm has the shortest runtime.

Researchers use four similarity measures: cosine similarity, Mean Squared Difference similarity, Pearson Correlation Coefficient, and Pearson Correlation Coefficient using baseline to compute the similarity between items in features

EVALUATION

At this stage, the system's effectiveness, efficiency, and performance are evaluated using the *Mean Absolute Error (MAE)* and *Root Mean Square Error (RMSE)* evaluation matrix. Researchers compared the effect of using the K-Nearest Neighbour with Means algorithm and K-Nearest Neighbour basic with four similarity measurements.

The evaluation results will be presented in the table and analyzed, the best modelling. The best modelling will be used in the deployment phase. Researchers can return to the Business Understanding stage to review whether the modelling has achieved the research objectives.

DEPLOYMENT

Deployment is carried out when the results at the evaluation stage are considered quite good. This stage uses the best modeling based on the previous evaluation stages. The results of modeling with test data become a dataset for deployment, so that the dataset contains the data from the modeling results. The data consists of 1431 test data, with 5 columns, namely *reviewID, asin, est, overall,* and *Details.* Then the data is used to provide recommendations with *asin* input. The recommendation system is designed to provide output in the form of some recommended products according to user preferences.

RESULT AND DISCUSSION

In this section, the authors will explain the result of this research. Experiments on a LENOVO core i3-6006U device with 4GB RAM on 64-bit Windows OS were performed in Python 3.9.0.

The researcher uses the *NumPy, pandas, seaborn,* and *surprise libraries*. Business understanding and data understanding steps go hand in hand so that the data used is appropriate and supports the achievement of the objectives of this study. To proceed to the next stage, these two stages must be reviewed repeatedly. The data is explored in number, with features in it. Initial research data with a size of 10261 has 27 blank lines of reviewerName and seven null values of reviewText. Feature *Overall* (Rating) the lowest value is 1, and the highest is 5.

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A2IBPI20UZIR0U	1384719342	cassandra tu "Yeah, well, that's just like, u...	[0, 0]	Not much to write about here, but it does exac...	5.0	good	1393545600	02 28, 2014
1	A14VAT5EAX3D9S	1384719342	Jake	[13, 14]	The product does exactly as it should and is q...	5.0	Jake	1363392000	03 16, 2013
2	A195EZSQDW3E21	1384719342	Rick Bennette "Rick Bennette"	[1, 1]	The primary job of this device is to block the...	5.0	It Does The Job Well	1377648000	08 28, 2013

Figure 3. Head data used

The total number of reviewers is 1429 users, and the number of products reviewed is 900. In this data, one user can review several items repeatedly, and similarly, 1 item can be reviewed multiple times either from the same person or different people. Figure 3 shows examples of data that grouping by users and the number of reviews given.

```
reviewerID
ADH0O8UVJOT10     42
A15TYOEWBQYF0X    38
A1L7M2JXN4EZCR    38
A2EZWZ8MBEDOLN    36
A2NYK9KWFMJV4Y    34
                  ..
A9KOC3PXLXYSB      5
A25T143MKB0K82     5
A9AETC0WEPZAM      5
A256QA9N8ZK520     5
A2Z15UQEUTE3T9     5
Name: overall, Length: 1429, dtype: int64
```

Figure 4. Results of data grouping

User with reviewerID ADH0O8UVJOT10 gave 42 reviews on the data we used, and in this data, 1 user reviewed at least 5 times. If the data is grouped by the number of items *(asin)* and the number of reviewers, the highest number of reviews for an item is in the range of 163 reviewers. These goods can be assumed to be products that have been sold for at least 163 purchases.

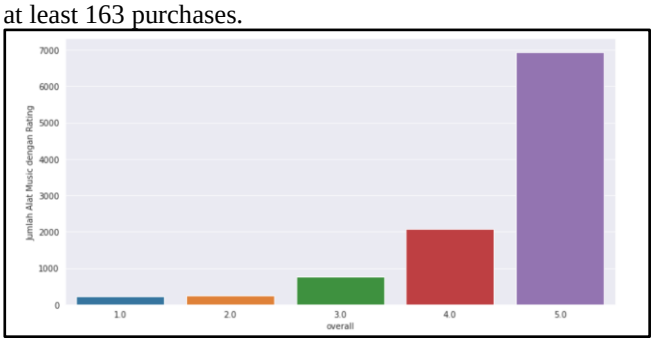

Figure 5. Chart Number of musical instruments (asia) and rating (overall)

Almost 7000 items rating of 5, about 2000 with a rating of 4, and less than 1000 items get a rating of 3 or less than 3.

Based on the understanding of the data used, to create a collaborative filtering recommendation system, the researchers chose to only use data of goods (*asin*) with at least 20 reviews (overall). This is enforced, at least the recommendations that will be given are at least items that have been reviewed enough, whether bad or not the value of the review. This selected data amounted to 3577 data, with an average review of each item as many as 91 reviews and the most reviews, namely 163 reviews. The assessment of the 3577 items has an average rating of 4,547, with a maximum value of 5, and the lowest rating of 3,807. The selected data is inserted into a new dataframe, and deletes some of the columns. After selecting the data, the data frame consists of 3577 items with 3 columns, namely *reviewerID, asin,* and *overall.* This data is used by researchers for modeling.

The modeling in this study begins with defining a rating scale, which is 1-5 and separating the datasets into train and test data. This research uses a ratio of 6:4, and 7:3 with a random state of 10. Thus, the train data used are 2146 and the test data is 1431 lines.

This test uses the KNN Basic and KNN with Means algorithms. The researcher also conducted modeling with k=5 and 4 similarity measurements, namely *Cosine Similarity (cosine), Mean Squared Difference (msd), Pearson Correlation Coefficient (pearson),* and *Pearson Correlation Coefficient - Baseline (pearson_baseline).*

The test results from this study can be seen in the table 2.1 below :

Table 2.1 Table of Result (70% Data Train)

	KNNBasic		KNNWithMeans	
	MAE	RMSE	MAE	RMSE
cosine	0,6030	0,9316	0,6101	0,9107
msd	0,634	0,9361	0,6108	0,9150
pearson	0,6341	0,8734	0,6257	0,8931
pearson_baseline	0,6238	0,9002	0,6144	0,8853

Table 2.2 Table of Result (60% Data Train)

	KNNBasic		KNNWithMeans	
	MAE	RMSE	MAE	RMSE
cosine	0,5998	0,9677	0,6182	0,9497
msd	0,6024	0,9697	0,6214	0,9525
pearson	0,6360	0,9010	0,6257	0,8931
pearson_baseline	0,6304	0,9430	0,6336	0,9327

The table above shows the comparison evaluation value of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) with similarity measurements using cosine, msd, Pearson, and pearson_baseline KNNBasic and KNNWithMeans. More smaller the MAE and RMSE values, more better the algorithm's performance.

The best results acording by the RMSE, calculation of the 70% data train models using KNNBasic algorithm with Pearson similarity measurements, and by the MAE with 60% data train using KNNBasic algorithm with cosine similarity.For the RMSE result was 0.8734, and for the MAE result was 0.5998. The recommendation system in this study is able to run according to the design.

CONCLUSION AND RECOMMENDATION

Collaborative Filtering using the K-Nearest Neighbor algorithm can be used in a musical instrument recommendation system.The RMSE (Root Mean Square Error) result obtained is 0.8734 for schema that divides dataset into 70% data train and 30% dataset using KNNWith Means with pearson measurements, and the MAE (Mean Absolute Error) result obtained is 0.5998 with schema 60% data train and 40% data test using KNNBasics algorithm with cosine similarity.

In this recommendation system, collaborative Filtering is used using the KNN algorithm to determine recommendations following user preferences. Other methods in the recommendation system, such as content-based Filtering or hybrid filtering methods, can be used to develop.

REFERENCE

Ahuja, R., Solanki, A., & Nayyar, A. (2019). Movie recommender system using k-means clustering and k-nearest neighbor. Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019,263–268. https://doi.org/10.1109/CONFLUENCE.2019.8776696

Arai, K., Kapoor, S., & Bhatia, R. (Eds.). (2020). Intelligent Computing. In Intelligent Computing Proceedings of the 2020 Computing Conference, (Vol. 1228). Springer International Publishing. https://doi.org/10.1007/978-3-030-52249-0

Baskota, A., & Ng, Y. K. (2018). A graduate school recommendation system using the multi-class support vector machine and KNN approaches. Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018, 277–284. https://doi.org/10.1109/IRI.2018.00050

Erlangga, E., & Sutrisno, H. (2020). Sistem Rekomendasi Beauty Shop Berbasis Collaborative Filtering. *EXPERT: Jurnal Manajemen Sistem Informasi Dan Teknologi*, *10*(2),47. https://doi.org/10.36448/jmsit.v10i2.1611

Gupta, M., Thakkar, A., Aashish, Gupta, V., & Rathore, D. P. S. (2020). Movie Recommender System Using Collaborative Filtering. Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, Icesc, 415–420. https://doi.org/10.1109/ICESC48915.2020.9155879

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Applying Collaborative Filtering to Usenet News. Communications of the ACM, 40(3), 77–87. https://doi.org/10.1145/245108.245126

Li, G., & Zhang, J. (2018). Music personalized recommendation system based on improved KNN algorithm. Proceedings of 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2018, Iaeac, 777–781. https://doi.org/10.1109/IAEAC.2018.8577483

Muneer, A., Fati, S. M., & Al-Ghobari, M. (2021). Location-Aware Personalized Traveler Recommender System (LAPTA) Using Collaborative Filtering KNN. Computers, Materials & Continua. https://doi.org/10.32604/cmc.2021.016348