ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Product Recommendation System Using Implicit Feedback Based on Collaborative Filtering in E-Commerce

Muhammad Nugraha Mahardhika[1], Fajar Rahayu[2*], Achmad Zuchriadi[3]
[123]Electrical Engineering, Faculty Of Engineering, Universitas Pembangunan Nasional Veteran Jakarta
Email: [1]dhikanugraha8@gmail.com, [2]fajarrahayu@upnvj.ac.id, [3]achmad.zp@upnvj.ac.id

Abstract

One of the marketing strategies of e-commerce companies is a recommendation system that is used to predict interesting product information based on the characteristics of each user. However, recommendation systems generally use explicit feedback as a value of user interest in a product, giving rise to data limitations (cold-start) problems because they are only based on transaction data that has been assessed by users. Another solution could be to use implicit feedback to avoid cold-start issues based on the number of user transactions for stores and product categories. In this research, the algorithm used is Singular Value Decomposition (SVD) to find similarities between one user and another user based on the feedback value. The model results show good performance with RMSE scores ± 0.865 and MAE ± 0.508.

Keywords*: Recommendation System, Cold-start, Implicit Feedback, Singular Value Decomposition (SVD).*

1. Introduction

The existence of e-commerce as an online buying and selling platform makes these activities much easier. E-commerce allows sales transactions to be carried out online so that buyers do not need come to the store. Thus, the sales transaction process will be able to save time and costs, buyers can also get new recommendations regarding the product to be purchased based on comparisons between user patterns [1]. It is not surprising that currently the growth of e-commerce in Indonesia is very high, resulting in massive scale transaction data that can be used as a marketing strategy by e-commerce companies to increase the number and value of transactions. One of the strategies currently being implemented is the Recommendation System, a tool for estimating interesting product information based on the suitability of each user's characteristics with the help of machine learning

In the recommendation system, there is a Collaborative Filtering approach as a means of filtering or evaluating items using other people. Collaborative filtering performs filtering of data based on the similarity of user characteristics so that it is able to provide new information to users because the system provides information based on the pattern of a group of users who are almost the same [2]. Generally, the collaborative filtering approach uses explicit feedback in processing the similarity pattern of the 'rating' given by the user to the product purchased. However, this causes a cold-start problem when there are new transactions, this information cannot be recommended to any user until they have been rated by someone [3]. Cold start is a condition when a new user has never given a rating on a product, so that the information obtained for the direction of interest from users is difficult to know [4]. Lack of information from new users, recommendation systems will have difficulty predicting and recommending a product to these users [5].

In an effort to solve cold-start problems that often occur, this research will use implicit feedback which allows the system to identify the user's interest in a product implicitly. This approach does not require a rating given by the user as a reference for sentiment towards the product, but instead uses the number of transactions for each user at a particular store and the number of transactions for the type of item category purchased. This is used as a theoretical concept because there is a tendency for a user to depend heavily on the store, and the information on the types of categories of product purchased can be used as recommendations that are more specific to each user.

So that through this method, any transaction data obtained can be directly used for the recommendation system without waiting for the rating given by the user. This information is then weighted as a label or dependent variable in the recommendation engine modeling. Furthermore, to evaluate a model, the ability test metrics used are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

2. Research Methods

The research methodology is the steps taken in carrying out the research. The first step that needs to be done is the stage of identifying what problems occur, namely the problem of data limitations. Then a data collection stage is needed using a transaction dataset in e-commerce to be processed using several techniques both in training and testing. These results will later be analyzed so that the system can produce product recommendations for each user.

2.1 Research Stages

There are several flow formulations or procedures carried out to achieve the research objectives shown in Fig 1.

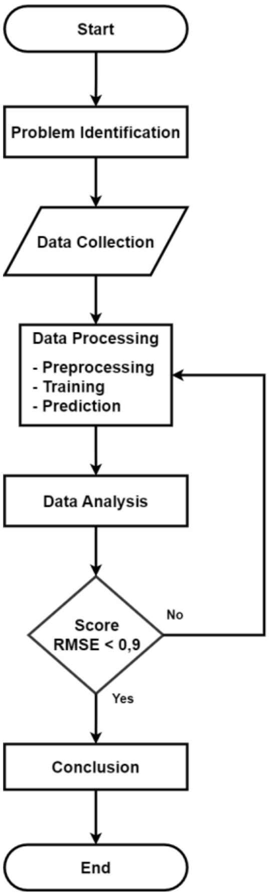


Figure 1. Research Flow

Based on the research flow shown in Figure 1, the first step is to identify what problems are occurring, namely the problem of data limitations (cold-start), which is currently still a polemic on recommendation systems based on collaborative filtering in e-commerce. To overcome these problems, it is necessary to carry out the data collection stage first using a dataset of buying and selling transactions in e-commerce. Then the next stage is processing data starting from pre-processing activities, data training, model predictions, and evaluation.

2.2. Problem Identification

Recommendation systems that work in e-commerce generally still constrained by cold-start problems (data limitations) because they still use explicit feedback such as ratings, stars, likes and thumbs up. Therefore, in this study the authors designed a recommendation system for e-commerce implicit feedback based on collaborative filtering including the number of transactions for each user in a store and the category of product purchased. The information from the two feedbacks will be combined to provide the best item recommendations. So with this design, the cold-start problem can be resolved.

2.3. Data Collection

This research requires transaction data on e-commerce with a large scale because it will later affect the performance of the model. Data taken using the e-commerce transaction dataset of PT. XYZ from July 2020 to February 2023. Dataset information and columns used are shown in Table 1.

Table 1. Description E-Commerce Transaction Dataset of PT. XYZ

Column	Unique Number	Data Type	Number of Dataset
userID	12752	String (Object)	462920 Rows
itemID	182507		
umkmID	6277		
category	46		

Information:

- 'user_id' is the identity number of the user or buyer
- 'item_id' is the identity number of the product
- 'umkm_id' is the identity number of the UMKM or seller
- 'main_category' is the main category type of the item

2.4. Training Techniques

In this study, the training technique used was using a method, namely Matrix Factorization. Matrix factorization is a way to make a matrix into two or more multiplication matrices. Suppose A is a matrix, then the factorization of A can be in the form A= A_1A_2 or A = $A_1A_2A_3$…, with adjusted sizes for A_i [6]. Matrix factorization is used in embedding the compatibility of features with one another which makes it possible to look for the relationship whether it is related or not with the aim of providing predictive results for the user's assessment of a product. The matrix factorization algorithm that will be used is Singular Value Decomposition (SVD). Mathematically, SVD is designed to make a matrix decomposed from Matrix M into 3 matrix parts as shown in Figure 2:

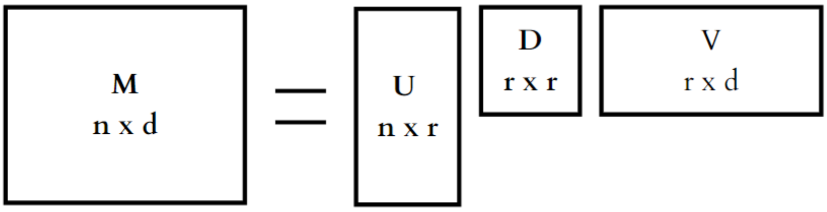


Figure 2. SVD Equation Formula

Information:

- Matrix U is a matrix of size n×r orthogonal matrix of eigenvectors $M^{\mathrm{T}}M$
- Matrix D is a matrix of size r×r which is a diagonal matrix that has singular values.
- Matrix V^T is a matrix of size r×d orthogonal matrix of eigenvectors MM^{T}

Matrix D is a diagonal matrix of non-negative numbers with the sum of *r* following the rank of matrix M. Matrix U and V have the exact column values so they have the same r value, that's why they are called left and right singular vectors. The matrix can be reduced to the power value k which can make the value of matrix D adjust to the power k and converted into the values of three factorization matrices namely n×k, k×k, dan k×d, with k<r. SVD provides a low k rating that retains the original M value by reducing the singular k value. It is called D_k matrix reduction because D is done by maintaining the D_k value. Matrix U and V are also reduced to produce Uk and Matrix V_k. Matrix U_k is obtained by removing $(r-k)$ column in matrix U while matrix V_k is obtained by removing $(r-k)$ rows of matrix V. When the three matrices are reduced, you will get matrix M_k [7].

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

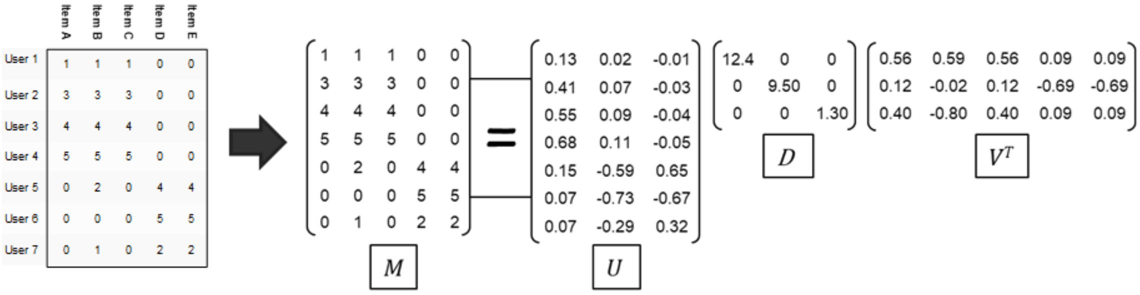


Figure 3. Matrix Transformation on SVD

In the SVD algorithm, each column will be mapped first as shown in Figure 3. Some of these columns consist of user identity (user id), product identity (item id), and feedback values (ratings) which are combined in one dataframe. An overview of the SVD algorithm in the parable, Matrix D can be said to be a bridge between two things, namely information about column U (users) and information about column V (products). From these results, SVD will look for user similarities in a group based on the number of entity values with a certain similarity value to fill in the feedback value that has not been given (rating value = 0) based on the predicted results.

2.5. Testing Techniques

To measure the accuracy of a system, an error value evaluation method is needed so that a model can run with optimal performance. Evaluate the error value using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) equations. Both are needed in calculating the average error value between the predicted results and the actual value. The smaller MAE and RMSE values, the better the system is made. However, if the MAE and RMSE are greater, then the system is not working optimally. Mean absolute error (MAE) is one of the accuracy metrics for continuous data. MAE shows the average absolute result of the difference in error values between the predicted results and the actual value. In the formula, MAE is stated as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Information:
- y is the actual/actual result value
- ŷ is the predicted/forecasting result value
- *i* is the order of data in the database
- *n* is the number of data

Meanwhile, Root Mean Square Error (RMSE) itself is an alternative method of evaluating forecasting techniques in measuring the accuracy of the prediction results of a model. The RMSE value is the root of the average error difference value. The mathematical formula for RMSE is presented in the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Information:
- y is the observed/actual value
- ŷ is the predicted/forecasting result value
- *i* is the order of data in the database
- *n* is the number of data

3. Results and Analysis

The results and analysis will show the stages by applying several predetermined methods. The data to be processed is then analyzed with several evaluation metrics to test how well the model performs.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

3.1. Preprocessing

Before entering the training stage, the data must go through the preprocessing stage first because this stage greatly influences the prediction results. This stage is very necessary because the data that has been collected needs to go through several processes before the data enters the training stage. All of these stages are carried out using tools to process data using the Python programming language, namely Google Colaboratory. Some of the python libraries used at this stage are Pandas, Matplotlib, Recommenders and Scikit-learn

1. Data Cleaning

   Basically, the dataset still dirty so it won't be able to enter into the training. Unstructured data such as nan values and disorganized are cleaned by dropping certain rows. In addition, duplicate data must also be removed so as not to interfere with the training process later. Therefore, the data cleaning process is needed to ensure that the data can be processed

2. Data Manipulation

   After going through the data cleaning stage, the data needs to be further modified to determine the labels needed later at the training stage. The label used is the number of transactions for the store and the number of product category transactions. The label will be manipulated as if it is a value of user interest in the feedback used. This stage begins with forming 2 dataframe that contain information about the value of each feedback. The first dataframe contains user transactions for each store, while the second dataframe contains user transactions for each item category.

3. Data Scaling

   The next stage is data scaling, which is needed to normalize and generalize the feedback value because it has a very large maximum and minimum range. Data scaling stage also makes the feedback value appear to be the user's rating of interest in certain variables with a range of 1 to 5. This will later facilitate understanding in the process of evaluating the performance of the machine learning model that will be created.

4. Data Splitting

   Before entering the training stage, the data needs to be divided first into training data and testing data with a training ratio of 80% and testing of 20%. This comparison is the most ideal division because apart following several previous studies, the result of performance also the most stable and optimal among other comparisons. This process is carried out twice considering that there are 2 dataframe with different feedback value information. Furthermore, the training data will be used in the training process while the testing data is used to evaluate the results of the performance of the model that has been made.

3.2. Data Training

Data that has gone through the preprocessing stage will be trained as a form of machine learning. With the scikit-surprise library, training is carried out using the Singular Value Decomposition (SVD) algorithm to find similarities between one user and another based on the feedback value given to the store and product category purchased. Because in the previous stage 2 dataframes with different information were created, training was carried out for each dataframe. The training process uses training data that has been split in the previous stage. At Table 2, there are several parameters that used during data training process.

Table 2. Data Training Parameters

Parameter	**Value**
reader(*rating_scale*)	(1, 5)
random_state	0
n_factors	200
n_epochs	30

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Based on Table 2, the "reader" parameter is needed to limit the minimum and maximum values of the prediction results in the form of a scale. The number of entities or groups created in the training process, requires the value of 'n_factor' to be given at 200. The number of iteration values for the 'n_epoch' parameter that is carried out is 30. The results of the data training process will form a model which will later be used to predict the value of each feedback user. Basically, parameter determination in training must be carried out by trial and error by changing random_state and n_factor values repeatedly to get a combination of parameter values with the best and optimal model performance results.

3.4. Evaluation

The evaluation stage begins with predicting the feedback value on the testing data using the model that has been created. The model prediction process is also carried out on 2 dataframes with different feedback value information. From the results of the predictions that have been made, how well the performance of the model will be evaluated. Because the feedback value is continuous (numeric), the evaluation process uses the MAE and RMSE metrics which calculate the average error value of the prediction results. Sedangkan jika nilai MAE dan RMSE yang tinggi justru merepresentasikan performa model yang kurang baik. The MAE and RMSE scores for the two prediction results are shown in the table below.

Table 3. Evaluation Metrics Model Performance

Metrics	Store Predictions	Category Prediction	Average
RMSE	0.836	0.893	0.865
MAE	0.491	0.525	0.508

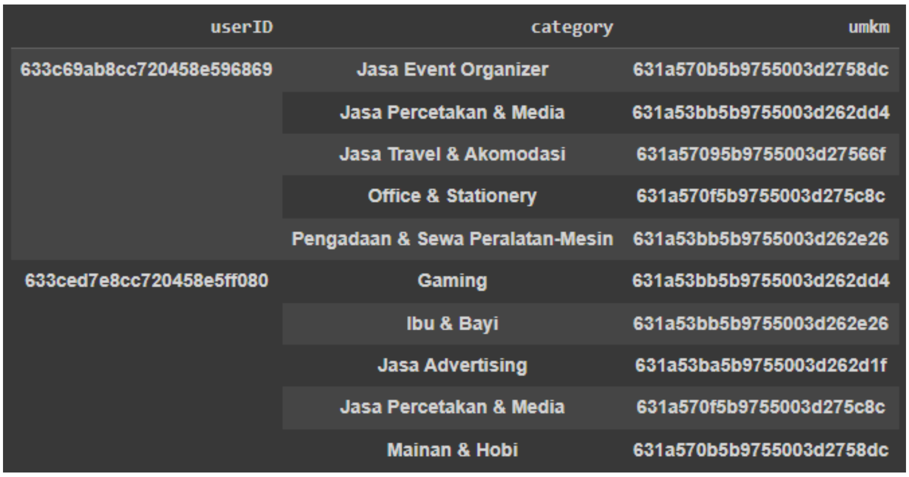
Based on Table 3, the model prediction results show a good score. These results also meet the desired target criteria, namely an RMSE value of less than 0.9 (Netflix Prize Competition). The RMSE value will always be greater than the MAE value. This is because the RMSE value uses the squared difference first before being rooted. So that the RMSE will increase very significantly if there is a very large difference between the predicted results and the testing data.

3.4. Overall Predictions

After going through the performance feasibility evaluation stage, the model will be used to predict the value of each user's feedback for all variables both for stores and product categories. Each user will be taken 5 prediction results with the highest feedback value. Overall prediction results will be displayed in the form of a dataframe based on the predicted feedback value of each.

3.5. Product Recommendations

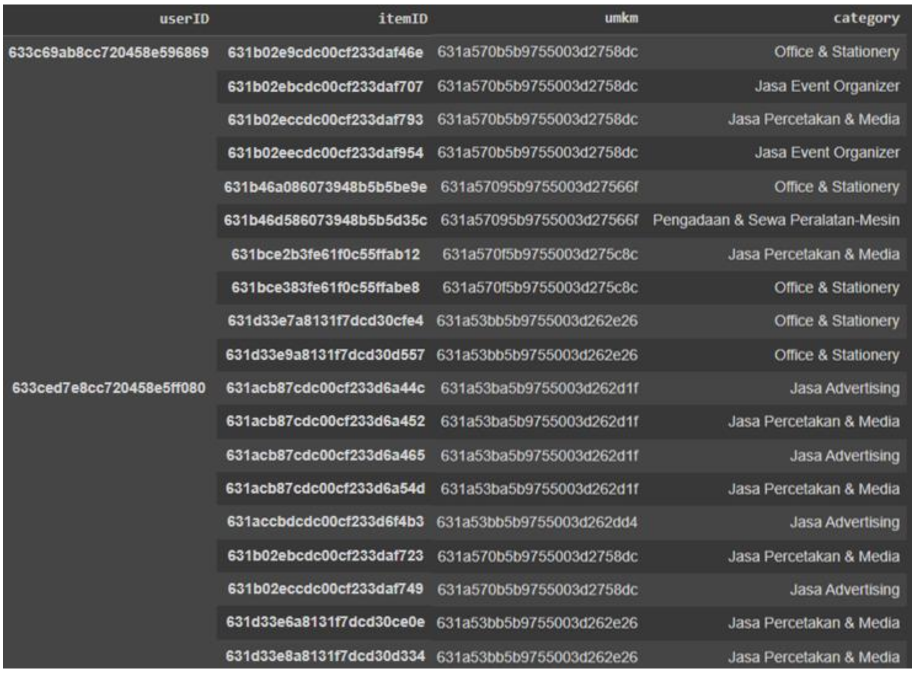
After obtaining the overall prediction results for each user for both types of feedback, namely stores and product categories, the next step is to provide product recommendations to each user. Because there are 2 prediction results that contain feedback value information, they need to be combined into one dataframe shown in the image below.



userID	category	umkm
633c69ab8cc720458e596869	Jasa Event Organizer	631a570b5b9755003d2758dc
	Jasa Percetakan & Media	631a53bb5b9755003d262dd4
	Jasa Travel & Akomodasi	631a57095b9755003d27566f
	Office & Stationery	631a570f5b9755003d275c8c
	Pengadaan & Sewa Peralatan-Mesin	631a53bb5b9755003d262e26
633ced7e8cc720458e5ff080	Gaming	631a53bb5b9755003d262dd4
	Ibu & Bayi	631a53bb5b9755003d262e26
	Jasa Advertising	631a53ba5b9755003d262d1f
	Jasa Percetakan & Media	631a570f5b9755003d275c8c
	Mainan & Hobi	631a570b5b9755003d2758dc

Figure 4. Results of Top 5 Prediction Types of Feedback

Based on Figure 4, it can be seen that each user has the top results of 5 store recommendations and 5 category recommendations. These results will be related like nodes that are connected to each other. If relationship between the store and the category meets the desired conditions, then product information related to the relationship between the two is obtained. However, if one user cannot fulfil the desired condition at all, then recommendations will be based on the products sold by that store at random. So that the final output of the system that has been created will produce 10 product recommendations for each user as shown in Figure 5.



userID	itemID	umkm	category
633c69ab8cc720458e596869	631b02e9cdc00cf233daf46e	631a570b5b9755003d2758dc	Office & Stationery
	631b02ebcdc00cf233daf707	631a570b5b9755003d2758dc	Jasa Event Organizer
	631b02eccdc00cf233daf793	631a570b5b9755003d2758dc	Jasa Percetakan & Media
	631b02eecdc00cf233daf954	631a570b5b9755003d2758dc	Jasa Event Organizer
	631b46a086073948b5b5be9e	631a57095b9755003d27566f	Office & Stationery
	631b46d586073948b5b5d35c	631a57095b9755003d27566f	Pengadaan & Sewa Peralatan-Mesin
	631bce2b3fe61f0c55ffab12	631a570f5b9755003d275c8c	Jasa Percetakan & Media
	631bce383fe61f0c55ffabe8	631a570f5b9755003d275c8c	Office & Stationery
	631d33e7a8131f7dcd30cfe4	631a53bb5b9755003d262e26	Office & Stationery
	631d33e9a8131f7dcd30d557	631a53bb5b9755003d262e26	Office & Stationery
633ced7e8cc720458e5ff080	631acb87cdc00cf233d6a44c	631a53ba5b9755003d262d1f	Jasa Advertising
	631acb87cdc00cf233d6a452	631a53ba5b9755003d262d1f	Jasa Percetakan & Media
	631acb87cdc00cf233d6a465	631a53ba5b9755003d262d1f	Jasa Advertising
	631acb87cdc00cf233d6a54d	631a53ba5b9755003d262d1f	Jasa Percetakan & Media
	631accbdcdc00cf233d6f4b3	631a53bb5b9755003d262dd4	Jasa Advertising
	631b02ebcdc00cf233daf723	631a570b5b9755003d2758dc	Jasa Percetakan & Media
	631b02eccdc00cf233daf749	631a570b5b9755003d2758dc	Jasa Advertising
	631d33e6a8131f7dcd30ce0e	631a53bb5b9755003d262e26	Jasa Percetakan & Media
	631d33e8a8131f7dcd30d334	631a53bb5b9755003d262e26	Jasa Percetakan & Media

Figure 5. Product Recommendation Results in Detail

4. Conclusion

The recommendation system is a solution that is implemented to be able to use implicit feedback because it does not require a rating given by the user, by only using the number of transactions for each user for the store and product category. So that all data can be included in training so that the recommendation system is able to work optimally to provide the best results. Designing a recommendation system using implicit feedback was successfully carried out by establishing several rules and adding modifications to overcome the problem of limited data. The algorithm used in this training is Singular Value Decomposition (SVD). The model performance obtained meets the desired criteria (RMSE < 0.9) with an RMSE value of ± 0.865 and an MAE value of ± 0.508.

Reference

[1] H. Februariyanti, A. D. Laksono, J. S. Wibowo and M. S. Utomo, "Implementasi Metode Collaborative Filtering Untuk Sistem Rekomendasi Penjualan Pada Toko Mebel," *Jurnal Khatulistiwa Informatika,* vol. IX, no. 1, 2021.

[2] A. Mulyana and S. Yuliyanti, "Aplikasi E-commerce Dengan Sistem Rekomendasi Berbasis Collaborative Filtering: Pada Toko Nocturnal," *Jurnal Teknologi Informasi dan Komunikasi,* vol. VII, no. 2, 2018.

[3] R. Burke, "Hybrid Web Recommender Systems," in *The Adaptive Web : Methods and Strategies of Web Personalization*, Heidelberg, Springer, 2007, pp. 377-408.

[4] B.-H. Huang and B.-R. Dai, "A Weighted Distance Similarity Model to Improve The Accuracy of Collaborative Recommender System," in *2015 16th IEEE International Conference on Mobile Data Management*, Pittsburgh, IEEE, 2015, pp. 104-109.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

[5] S. Sylvia and S. Lestari, "Implementasi K-Means Dalam Mengatasi Masalah Cold Star Pada Collaborative Filtering," in *Prosiding Seminar Nasional Darmajaya*, Yogyakarta, 2002.

[6] N. Nurmalasari, Y. Yanita and I. M. Arnawa, "Faktorisasi Matriks," *Jurnal Matematika UNAND,* vol. VIII, no. 1, pp. 242-248, 2019.

[7] C. Wibisono, L. S. Haryadi, J. E. Widyaya and S. L. Liliawati, "Sistem Rekomendasi Suku Cadang Berdasarkan Item Based Filtering," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. VII, no. 1, pp. 10-19, 2021.