Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 1 (2023) 1285 - 1291 ISSN Media Electronic: 2580-0760

Imputation Missing Value to Overcome Sparsity Problems in The Recommendation System

Sri Lestari[1], M. Elrico Afdila[2], Yan Aditiya Pratama[3]
[1,2]Faculty of Computer Science, Institute of Informatics and Business Darmajaya, Bandar Lampung, Indonesia
[3]Faculty of Business Economics, Institute of Informatics and Business Darmajaya, Bandar Lampung, Indonesia
[1]srilestari@darmajaya.ac.id, 1911010061, [2]1911010061@mail.darmajaya.ac.id, [3]yanaditiyapratama@darmajaya.ac.id

Abstract

A recommendation system is a system that provides suggestions or recommendations for a product or service for its users. One of the problems encountered in the recommendation system is sparsity, namely the lack of available data for analysis, resulting in poor performance of the recommendation system because it cannot provide the proper recommendations. On this basis, this study proposes the mean method and the stochastic Hot-Deck Method to calculate missing values to improve the quality of the recommendations. The experimental results show that the hot-deck imputation method gives better results than the mean imputation method with smaller RMSE and MAE values, namely 2,706 and 2,691.

Keywords: missing value; stochastic hot-deck; imputation

1. Introduction

The recommendation system consists of several methods such as content-based, collaborative filtering, demographic information, and hybrid. Collaborative Filtering (CF) uses a database of consumer preferences to predict additional topics or products that consumers might also like. In many CF scenarios, there will be a list of m users {u1,u2,…, um} and a list of n items {i1,i2,…, in}, and each user, UI, has a list of items, Iui, that the user has rated, or from preferences inferred from their behaviour. Ratings can be taken from explicit indications, such as a scale of 1-5, or implicit indications, such as purchases or elections [1].

Collaborative filtering is the most successful method in recommendation systems. Collaborative filtering is divided into two categories, namely memory-based and model-based [2]. The memory-based approach uses patterns of similarity between users, often called user-based, and between services based on historical data, called item-based [3]. Item-based filtering works by collecting the attributes contained in an item and trying to find other items that have the same attributes, where similar items will be recommended to users [4]. However, CF has some drawbacks such as sparsity. Data similarity only takes common values, causing it to be unreliable if the data is sparse. Missing data will cause inaccurate parameter estimation due to reduced data size [5].

Data Sparsity is the occurrence of vacancies in the user-item data matrix, which is caused by the user rating in a small number of the number of items available in the database [6]. Rubin (2002) defines missing data based on three loss mechanisms: Missing Completely At Random (MCAR) when the probability of a case having an error value for the variable does not depend on a known value or missing data; Missing At Random (MAR) when the probability of a case having a missing value for a variable can depend on the known value but not on the value of the missing data itself; Missing Not At Random (MNAR) when the probability of an instance having a missing value for a variable can depend on the value of that variable [7].

Data sparsity is a problem that arises in various situations, such as when there are new users who are just using the service, it will be difficult to find the same preferences because there is not much information [8]. The latest films cannot be recommended until there are users who recommend them, and not all users will give good recommendations due to the lack of historical data for their recommendations. This can reduce the effectiveness of recommendation services that rely on the comparison of user recommendations so that predictions can be issued. From these problems, additional algorithms are needed to minimize data gaps in the recommendation system [9].

Accepted: 24-07-2023 | Received in revised: 10-11-2023 | Published: 26-11-2023

Imputation is one of the algorithms that can be used to solve the sparsity problem. One of the imputation methods used is to replace missing data with an average value or with a mode depending on the type of data. For numerical data, it is used to replace missing data with an average value, while for categorical data, it is used to replace missing data with the closest value.

The imputation process can also be described using linear regression and taking the imputed values as a random sample from a normal distribution. This is problematic if the residuals from the regression are not normally distributed (eg.: if the data is skewed) or if the relationship between the variables is nonlinear (eg.: height and age). For example, a variable that can only have positive values (for example, amount) can have negative imputed values. One option to overcome this problem is to transform the variables before imputation so that the transformed variables have a distribution that is closer to the normal distribution. For example, a logarithmic transformation, when applied to a positively skewed distribution, can produce a distribution that is closer to the normal distribution. As a final step, we can back-transform the imputation to return it to its original scale [10].

Another imputation method is Hot-deck. Hot-deck Imputation generally refers to Sequential Hot-deck Imputation, which means that the data set is sorted and missing values are imputed sequentially walking through observation after observation. Sorting data using predictor variables is selected based on its relationship with the variable to be imputed [11]. Stochastic Hot-deck imputation is a Hot-Deck Method that involves randomly selecting a donor record from complete cases in the data and using it to fill in the missing values for an incomplete case. Imputation is carried out stochastically, which means that the selection of donor records is random and depends on the distribution of donor records [12]. Currently, methods with better theoretical properties are available, but the Hot-deck Imputation method is still quite popular due to its simplicity and speed.

Several studies have been carried out to solve sparsity problems using imputation methods such as average values, closest values, Hot decks, and others. Meanwhile, in this research, we compared the Mean and Hot Deck impact methods with a movie dataset to find out which method performed better.

2. Research Methods

Research methodology uses a series of procedures to obtain data, analyze data, and draw conclusions based on the results of the analysis. The following is an explanation of the research methodology that can be carried out for missing value imputation research using the MovieLens 100K dataset.

The dataset used in this research is MovieLens 100K data, and Data Wrangling is the method used for cleaning, preprocessing, and transforming data [13]. Data wrangling is the process of transforming raw data into a usable form. This process can also be referred to as data processing or data repair. Usually, research will go through a data wrangling process before conducting data analysis to ensure data is reliable and complete [14]. The MovieLens 100K dataset can be downloaded for free from the official MovieLens website, As seen in Figure 1. This dataset contains 100,000 movie ratings by MovieLens users and consists of three files: a rating file, a movie information file, and a user file.

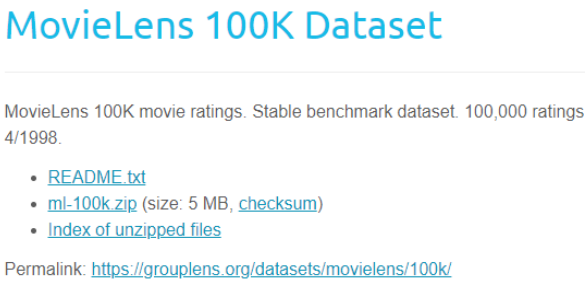

Figure 1. Movielens dataset information

Systematic sampling is a random sampling method that involves selecting sample units at certain intervals from the desired target population [15]. This method is carried out by determining the interval between units taken randomly from the population, and then choosing the first unit at random.

Systematic sampling is preferred over simple random sample selection when the risk of data manipulation is low. If the risk is high, where a researcher can manipulate the length of the interval to get the desired result, then a simple random sample selection technique would be more appropriate. [16].

This study uses systematic sampling on a population of 1000 users, with 10 intervals, so every 10 users will be taken as a sample, for example, the 10th, 20th, 30th users, and so on. In systematic sampling, the first sample unit is taken randomly, and then the interval is determined according to the desired number of sample units, and then the data will be classified. Classification is a form of data analysis that extracts models that describe data classes [17].

Data Imputation is carried out on sampling data, and then an analysis is carried out by comparing the results of imputation with one method with another. The following is an example of testing with the manual calculation imputation method on a small dataset of film ratings in Table 1.

In Table 1, the results are not perfect because they still have missing data. Then, it will be imputed with the mean method and the Hot-Deck Method.

DOI: https://doi.org/10.29207/resti.v7i6.5300

Table 1. Sample of Dataset with Missing Data

ID User	Age	Gender	Work	Movie 1	Movie 2	Movie 3	Movie 4
136	51	Male	Other	4	5	0	5
137	50	Male	Teacher	4	4	3	0
138	46	Female	Doctor	4	2	2	0
139	20	Male	Student	2	3	0	3
140	30	Female	Student	0	0	0	0

Imputation with the Mean method has the disadvantage of reducing the variance of the variables because the values entered are the same for each variable [18]. The initial calculation can be seen in Formula 1.

$$pref_{u,g} = \frac{\sum i \in I_g r_{ui}}{\|I_f\|} \quad (1)$$

$pref$u,g is the rating value from user u to genre g, $r$ui is the user u's rating to item i, Ig is the a collection of items that have a genre g.

The mean imputed value to be used in filling in the rating of an item is the mean value of the user's rating of the genres that the item belongs to [19]. The mean imputation calculation is in Formula 2.

$$impmean_{u,i} = \frac{\sum g \in G_i pref_{u,g}}{\|G_i\|} \quad (2)$$

$impmeanu$, I is the mean imputed value of user u for the item i, $pref$u,g is the user rating of the genre g, Gi is a collection of items to which the item belongs i.

Missing data on film 1 is filled with the average of all known scores in film 1:

(4+4+4+2)/4=3,5

Then the results follow the data containing integers to = 4. The imputation results obtained are then filled into all blank scores in film 1. Imputation is continued for scores in films 2, 3, 4 and so on. Table 2 is the mean imputed result. Transformation

Table 2. Mean Imputation Results

ID User	Age	Gender	Work	Movie 1	Movie 2	Movie 3	Movie 4
136	51	Male	Other	4	5	3	5
137	50	Male	Teacher	4	4	3	4
138	46	Female	Doctor	4	2	2	4
139	20	Male	Student	2	3	3	3
140	30	Female	Student	4	4	3	4

Different from the mean, Hot-deck uses other indicators as an average reference. Hot-deck Imputation generally refers to Sequential Hot-deck Imputation, which means that the data set is sorted and missing values are imputed sequentially walking through observation after observation [20]. When referring to stochastic by gross, the stochastic process is the set of random variables t{X(t),t∈T}. All possible values that can occur in the random variable X(t) are called the state space. One t value of T is called the index or time parameter. With this time parameter, the stochastic process can be divided into two forms, namely, first, if T = {0,1,2,3,…} then this stochastic process has discrete parameters and is usually abbreviated with the notation {X12}. (4). Second If T = {t | t ≥0} then the stochastic process has continuous parameters and is expressed by the notation {X(t) | t≥0}. (5)

Furthermore, this study uses gender parameters. As seen in film 1, the average female user has a score of 4, so the result of ID 140 users who are female is 4. Meanwhile, in film 3, the average male user has a score of 3, which causes a score of user IDs 136 and 139 to be 3. Imputation continues on the score in the film that has not been filled in and so on. The results of the Hot-deck imputation can be seen in Table 3.

The imputation process was carried out using the Mean and Hot-deck, and then it was followed by a comparative analysis using RMSE and MAE evaluations. RMSE and MAE are two evaluation metrics used to measure how close the imputed data is to the original missing data. In this analysis, the performance of the two methods is measured by calculating the RMSE and MAE from the resulting imputed data.

Table 3. Hot-deck Imputation Results

ID User	Age	Gender	Work	Movie 1	Movie 2	Movie 3	Movie 4
136	51	Male	Other	4	5	3	5
137	50	Male	Teacher	4	4	3	4
138	46	Female	Doctor	4	2	2	1
139	20	Male	Student	2	3	3	3
140	30	Female	Student	4	4	2	1

DOI: https://doi.org/10.29207/resti.v7i6.5300

In the previous dataset example, RMSE provides information about how far the model's average prediction is from the initial dataset in the same units as the imputed dataset value. The measurement of accuracy used in this study is RMSE (Root Mean Square Error) where RMSE calculates the difference and error values found between actual and forecast data. The RMSE value indicates the level of accuracy of the model being built. The smaller the RMSE value, the resulting accuracy will be higher [21], [22]. RMSE calculation using Formula 3.

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - y_i)^2} \quad (3)$$

RMSE = √(((5-5)^2+(0-4)^2+(0-4)^2+(3-3)^2+(0-4)^2)/5)

Through calculations that follow Formula 3 [23], it can be seen that the RMSE obtained is 3.09

MAE provides information about how far the model's average prediction is from the initial dataset in the same units as the MAE value. The lower the MAE value, the better the model performance [24]. Calculating MAE using Formula 4.

$$MAR = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (4)$$

MAE = ((5-5)+(0-4)+(0-4)+(3-3)+(0-4))/5

Through the calculations that follow Formula 3 [25], it can be seen that the MAE obtained is 2.4.

3. Results and Discussions

This research uses a dataset from GoupLens.org, namely 100k data consisting of 1682 movies, 943 users, and 100,000 ratings, so the data contains 93.7% sparsity. In addition, there is user demographic information, namely age, gender, occupation, and zip. (https://grouplens.org/datasets/movielens/100k/).

The initial step taken in this research was to download the Movielens 100K dataset with the. CSV (comma-separated values) extension and import it into a Google Colab notebook, as shown in Figures 2 and 3. This was done to facilitate the next step, namely data preprocessing.

The next stage is cleaning by conducting systematic sampling on the dataset by filtering the data. The data is taken starting from user 5 and continues with interval 5, the selected data movies start from movies 1 to movies 100. The sampling results can be seen in Figure 4.

After cleaning the data, the next step is data transformation. In the dataset, the missing data is still an integer namely the value 0, so the data type cannot be imputed. Then the data must be changed to empty data (NaN) so that it can be identified that the data is missing data. The results of the transformation can be seen in Figure 5.

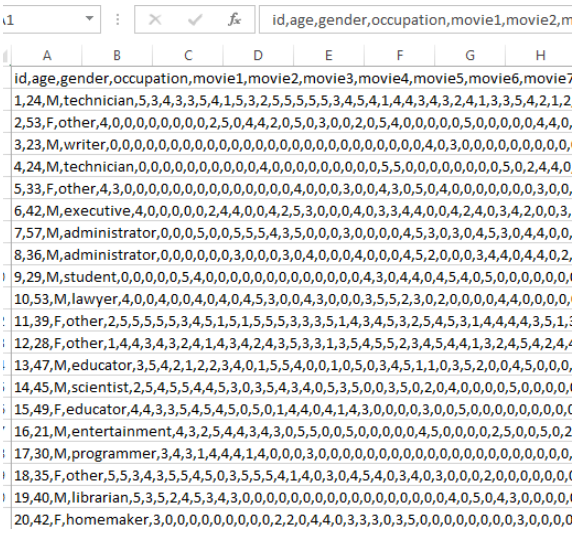

Figure 2. Raw Data

```
from google.colab import files

uploaded = files.upload()
```

Figure 3. Import Dependency

	id	age	gender	occupation	movie1	movie2	movie3	movie4	movie5	movie6	...	movie91	movie92	movie93
0	5	33	F	other	4	3	0	0	0	0	...	2	0	4
1	10	53	M	lawyer	4	0	0	4	0	0	...	4	0	4
2	15	49	F	educator	4	4	3	3	5	4	...	0	0	0
3	20	42	F	homemaker	3	0	0	0	0	0	...	5	0	0
4	25	39	M	engineer	5	0	0	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	480	57	M	retired	0	0	0	0	0	0	...	0	0	0
96	485	44	F	educator	0	0	0	0	0	3	...	0	0	0
97	490	29	F	artist	3	0	0	0	0	0	...	0	0	4
98	495	29	M	engineer	4	2	4	2	4	5	...	2	0	0
99	500	28	M	administrator	4	0	4	0	0	0	...	0	0	4

Figure 4. Sampling Result

```
df_NaN = df_movie.replace(0, np.NaN)
df_NaN
```

	movie1	movie2	movie3	movie4	movie5	movie6	movie7	movie8	movie9	movie10	...	movie91	movie92	movie93
0	4.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2.0	NaN	4.0
1	4.0	NaN	NaN	4.0	NaN	NaN	4.0	NaN	4.0	NaN	...	4.0	NaN	4.0
2	4.0	4.0	3.0	3.0	5.0	4.0	5.0	4.0	5.0	NaN	...	NaN	NaN	NaN
3	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	5.0	NaN	NaN
4	5.0	NaN	NaN	NaN	NaN	NaN	4.0	4.0	NaN	NaN	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5.0	NaN	NaN	...	NaN	NaN	NaN
96	NaN	NaN	NaN	NaN	NaN	3.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
97	3.0	NaN	NaN	NaN	NaN	NaN	3.0	NaN	4.0	NaN	...	NaN	NaN	4.0
98	4.0	2.0	4.0	2.0	4.0	5.0	4.0	NaN	5.0	NaN	...	2.0	NaN	NaN
99	4.0	NaN	4.0	NaN	NaN	NaN	5.0	4.0	4.0	3.0	...	NaN	NaN	4.0

Figure 5. Data Transformation

The data through the stages of cleaning, preprocessing and transforming indicate that the data is ready to be processed. Then it is continuous to do the imputation using the Mean and the Hot-Deck Method. For imputation the empty data mean will be imputed by the overall average of the data in each column, and then the results of the imputation will be rounded up to the nearest number. The mean imputation results can be seen in Figure 6.

DOI: https://doi.org/10.29207/resti.v7i6.5300

```
df_mean = df_NaN.fillna(df_NaN.mean())
df_mean
```

	movie1	movie2	movie3	movie4	movie5	movie6	movie7	movie8	movie9	movie10	...	movie91	movie92	movie93
0	4.000000	3.000000	3.26087	3.333333	3.4	4.28	3.688889	3.90625	3.685714	4.25	...	2.000000	3.9	4.000000
1	4.000000	3.185185	3.26087	4.000000	3.4	4.28	4.000000	3.90625	4.000000	4.25	...	4.000000	3.9	4.000000
2	4.000000	4.000000	3.00000	3.000000	5.0	4.00	5.000000	4.00000	5.000000	4.25	...	3.482759	3.9	3.894737
3	3.000000	3.185185	3.26087	3.333333	3.4	4.28	3.688889	3.90625	3.685714	4.25	...	5.000000	3.9	3.894737
4	5.000000	3.185185	3.26087	3.333333	3.4	4.28	4.000000	4.00000	3.685714	4.25	...	3.482759	3.9	3.894737
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	3.821429	3.185185	3.26087	3.333333	3.4	4.28	3.688889	5.00000	3.685714	4.25	...	3.482759	3.9	3.894737
96	3.821429	3.185185	3.26087	3.333333	3.4	3.00	3.688889	3.90625	3.685714	4.25	...	3.482759	3.9	3.894737
97	3.000000	3.185185	3.26087	3.333333	3.4	4.28	3.000000	3.90625	4.000000	4.25	...	3.482759	3.9	4.000000
98	4.000000	2.000000	4.00000	2.000000	4.0	5.00	4.000000	3.90625	5.000000	4.25	...	2.000000	3.9	3.894737
99	4.000000	3.185185	4.00000	3.333333	3.4	4.28	5.000000	4.00000	4.000000	3.00	...	3.482759	3.9	4.000000

Figure 6. Results of Mean Imputation on the Movie Dataset

In the next step, the imputation was carried out using the Hot-Deck Method with Hot-deck stochastics, using gender as an observation, missing data on a variable in the same data row will be matched with available data in other data rows that have the same gender value. Thus, the missing value in the variable being searched for will be filled in with the same value in that variable in the matching data row.

Data with the first enters the empty matrix, it is still missing and it will be filled with values by random division. Hot-deck imputation is carried out twice with different genders so that the values with male and female genders that are processed do not merge, then after the data is imputed, the two will be combined, as seen in Figure 7.

	gender	movie1	movie2	movie3	movie4	movie5	movie6	movie7	movie8	movie9	...	movie91	movie92	movie93
0	F	4.0	3.0	3.0	3.0	4.0	4.0	3.0	4.0	3.0	...	2.0	1.0	4.0
1	M	4.0	3.0	3.0	4.0	4.0	3.0	4.0	3.0	4.0	...	4.0	3.0	4.0
2	F	4.0	4.0	3.0	3.0	5.0	4.0	5.0	4.0	5.0	...	2.0	1.0	5.0
3	F	3.0	5.0	4.0	5.0	3.0	4.0	3.0	4.0	4.0	...	5.0	1.0	4.0
4	M	5.0	3.0	5.0	1.0	2.0	4.0	4.0	4.0	2.0	...	3.0	4.0	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	M	2.0	2.0	1.0	4.0	4.0	5.0	4.0	5.0	5.0	...	5.0	5.0	5.0
96	F	4.0	2.0	3.0	3.0	3.0	3.0	5.0	3.0	5.0	...	2.0	3.0	4.0
97	F	3.0	1.0	4.0	4.0	4.0	4.0	3.0	4.0	4.0	...	5.0	1.0	4.0
98	M	4.0	2.0	4.0	2.0	4.0	5.0	4.0	4.0	5.0	...	2.0	4.0	4.0
99	M	4.0	4.0	4.0	3.0	2.0	5.0	5.0	4.0	4.0	...	1.0	4.0	4.0

Figure 7. Hot-deck Imputation Results

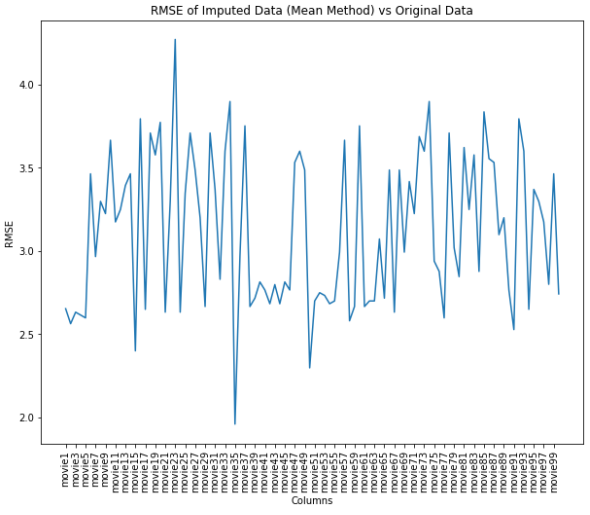

Figure 8. The Mean Imputed RMSE Graph

The next step is to evaluate the results of the imputation of the Mean and the Hot-Deck Method using the RMSE and MAE evaluations. RMSE and MAE are two evaluation metrics used to measure how close the imputed data is to the original missing data. In this analysis, the performance of the two methods is measured by calculating the RMSE and MAE from the resulting imputed data. Evaluation results can be seen in Figures 8 and 9 for the Mean method, and Figures 10 and 11 for the Hot-Deck Method.

```
# Total of mean RMSE result
print("Mean RMSE:", np.mean(rmse_scores))

Mean RMSE: 3.12069665323677
```

Figure 9. The Mean Imputed RMSE Results

So, it is known that the results of the RMSE (Root Mean Square Error) from the mean imputation in this study are: 3.120

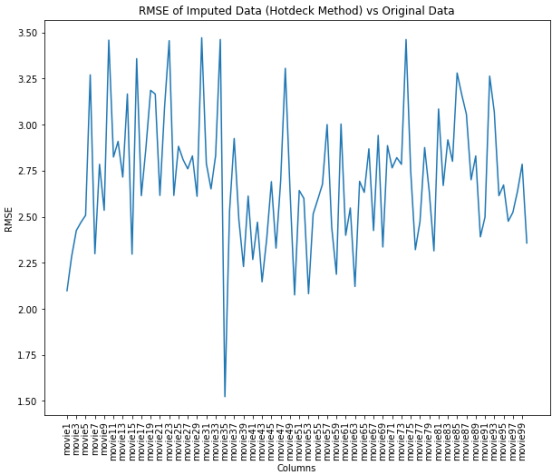

Figure 10. RMSE Graph of Imputed Hot-deck

```
# Total of Hot-deck RMSE result
print("Hot-deck RMSE:",
np.mean(rmse scores))

Hot-deck RMSE: 2.7064309707581113
```

Figure 11. RMSE of Hot-deck Imputation

The RMSE (Root Mean Squared Error) result from the Hot-deck imputation is 2,706

The next evaluation is MAE with the results shown in Figures 12 and 13 for the Mean method, and Figures 14 and 15 for the Hot-Deck Method.

DOI: https://doi.org/10.29207/resti.v7i6.5300

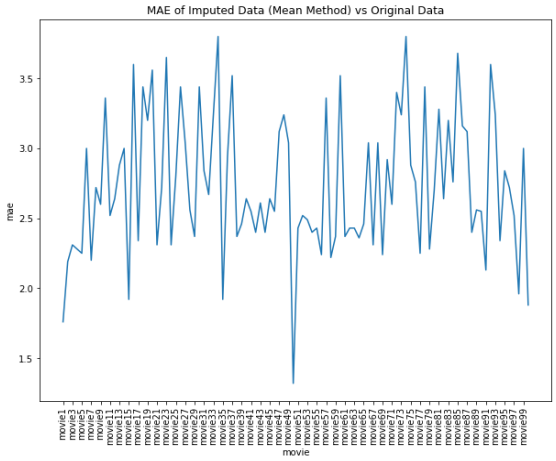

Figure 12. MAE of Imputed Mean Graph

```
# Total of mean MAE result
print("Mean MAE:", np.mean(mae))

Mean MAE: 2.7318999999999996
```

Figure 13. The mean imputed MAE results

Thus, it is known that the result of the MAE (Mean Absolute error) from the mean imputation in this study is 2,731

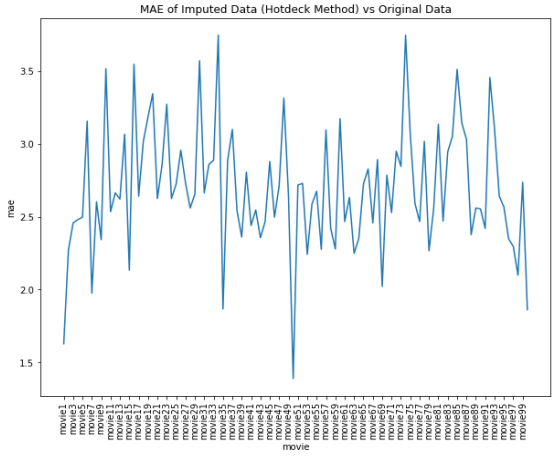

Figure 14. MAE Graph of Imputed Hot-deck

```
# Total of Hot-deck MAE result
print("Hot-deck MAE:", np.mean(mae))

Hot-deck MAE: 2.7451198206344163
```

Figure 15 MAE Results of Imputed Hot-deck

The imputation technique is used to overcome the sparsity problem in the recommendation system, and this study uses two imputation methods, namely Mean and Hot-deck. The imputation results were then evaluated using RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). The evaluation results can be seen in Table 4.

Table 4. Evaluation Results

Imputation methods	Evaluation RMSE	MAE
Mean	3.120	2.731
Hot-deck	2.706	2.691

The evaluation results show that there is a difference in value between the Mean and Hot-deck imputation techniques, namely, the RMSE value is 0.414 and the MAE value is 0.040, where the Hot-deck evaluation value is smaller in both the RMSE and MAE values. This is because it is influenced by the observation data factor, namely gender. The smaller the RMSE and MAE values indicate better performance in Hot-deck imputation.

4. Conclusion

Experimental results show that the imputation method can solve the sparsity problem and improve the quality of recommendations. This can be seen from the evaluation results of the Mean and the Hot-Deck Method with RMSE values of 3,120 and 2,706, respectively, and MAE values of 2,731 and 2,691. So, the Hot-deck imputation method gives better results than the Mean imputation method. That is because the smaller the RMSE or MAE, the better the performance of the imputation method in solving sparsity problems. Apart from that, it will produce higher-quality recommendations that are in line with user interests.

References

[1] Anang Furkon Rifai, Erwin Budi Setiawan, (2022). Memory-based Collaborative Filtering on Twitter Using Support Vector Machine Classification. *JURNAL RESTI*, 702 - 209.

[2] X. Wang, Z. Dai, H. Li, and J. Yang,(2020). A New Collaborative Filtering Recommendation Method Based on Transductive SVM and Active Learning. Discreet. Dyn. Nat. Soc., vol. 2020, no. 1.

[3] L. Ren and W. Wang,(2019). An SVM-based collaborative filtering approach for Top-N web services recommendation. Futur. Gener. Comput. Syst., vol. 78, pp. 531–543.

[4] von Hippel PT,(2020). How many imputations do you need? A two-stage calculation using a quadratic rule. Social methods res., 2020;49:699-718.

[5] Melinda, Imam Muttaqin, M., Nurdin, Y., & Bahri, A. (2023). Implementation of Word Recommendation System Using Hybrid Method for Speed Typing Website. *JURNAL RESTI*, 7(1), 7–14.

[6] Islamiyah, M., Subekti, P., Dwi Andini, T., & Asia Malang, S. (2019). Pemanfaatan Metode Item Based Collaborative Filtering Untuk Rekomendasi Wisata Di Kabupaten Malang. *Jurnal Ilmiah Teknologi Informasi Asia*, 13(2).

[7] G. Vink, "Roderick J. Little and Donald B. Rubin: Statistical Analysis with Missing Data," Psychometrika, 2002

[8] Subagyo, I., Dwi Yulianto, L., Permadi, W., Dewantara, A. W., & Hartanto, A. D. (2019). Sentiment Analisis Review Film Di IMDB Menggunakan Algoritma SVM Sentiment Analysis of Film Review at IMDB using SVM algorithm., *INFORMASI* (Vol. 47).

[9] H. Tahmasebi, R. Ravanmehr, and R. Mohamadrezaei (2021). Social movie recommender system based on deep autoencoder

DOI: https://doi.org/10.29207/resti.v7i6.5300

network using Twitter data. Neural Comput. Appl., vol. 33, no. 5, pp. 1607–1623.

[10] Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Canadian Journal of Cardiology, 37(9), 1322–1331.

[11] Raudhatunnisa, T., & Wilantika, N. (2022). Performance Comparison of Hot-Deck Imputation, K-Nearest Neighbor Imputation, and Predictive Mean Matching in Missing Value Handling, Case Study: March 2019 SUSENAS Kor Dataset. Proceedings of The International Conference on Data Science and Official Statistics, 2021(1), 753–770.

[12] Evenson, R. E. (2017). A Stochastic Model of Applied Research Author (s): Robert E . Evenson and Yoav Kislev Source: Journal of Political Economy, Vol . 84, No . 2 (Apr ., 1976), pp . 265-282

[13] Ritvik Voleti. (2020). Data Wrangling- A Goliath of Data Industry. *International Journal of Engineering Research And*, *V9*(08).

[14] Jiang, S., Kahn, J(2020). Data wrangling practices and collaborative interactions with aggregated data. Intern. J. Comput.-Support. Collab. Learn 15, 257–281.

[15] P. Wibowo and C. Fatichah(2021). In-depth performance analysis of the oversampling techniques for high-class imbalanced datasets. vol. 7, no. January, pp. 63–71.

[16] Xu, Y., Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. J. Anal. Test. 2, 249–262.

[17] Ilham, A. (2020). Hybrid Metode Boostrap Dan Teknik Imputasi Pada Metode C4-5 Untuk Prediksi Penyakit Ginjal Kronis. *Statistika*, *8*(1), 43–51.

[18] Dhimas Irnawan, F., Hidayah, I., & Nugroho, L. E. (2021). Metode Imputasi pada Data Debit Daerah Aliran Sungai Opak, Provinsi DI Yogyakarta. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, (Vol. 10).

[19] Ilham, A. (2020). Hybrid Metode Boostrap Dan Teknik Imputasi Pada Metode C4-5 Untuk Prediksi Penyakit Ginjal Kronis. Statistika, 8(1), 43–51.

[20] Fadillah, I. J., Muchlisoh, S., Statistika, P., & Stis, P. S. (2021). Perbandingan Metode Hot-Deck Imputation dan Metode KNNI dalam Mengatasi Missing Values. *Jurnal Ilmiah Politeknik Statistika STIS*, 275-285.

[21] Zahara, S., & Sugianto. (2021). Peramalan Data Indeks Harga Konsumen Berbasis Time Series Multivariate Menggunakan Deep Learning. *JURNAL RESTI*, 5(1), 24–30.

[22] Ilham, A. (2020). Hybrid Metode Boostrap Dan Teknik Imputasi Pada Metode C4-5 Untuk Prediksi Penyakit Ginjal Kronis. *Statistika*, *8*(1), 43–51.

[23] Moch Farryz Rizkilloh and Sri Widiyanesti (2022) .Prediksi Harga Cryptocurrency Menggunakan Algoritma Long Short Term Memory (LSTM). J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 6, no. 1.

[24] Azam Zamhuri Fuadi, Irsyad Nashirul Haq and Edi Leksono (2021). Support Vector Machine to Predict Electricity Consumption in the Energy Management Laboratory. *JURNAL RESTI*, Vol. 5 No. 3 466 - 473.

[25] I. M. Yudha Arya Dala, I. K. Gede Darma Putra, and P. Wira Buana (2021). Forecasting Cases of Dengue Hemorrhagic Fever Using the Backpropagation, Gaussians and Support-Vector Machine Methods. *JURNAL RESTI*, vol. 5, no. 2.

DOI: https://doi.org/10.29207/resti.v7i6.5300