RESWARA: Jurnal Riset Ilmu Teknik
2026, Vol. 4, No. 3, pp. 1056-1069
E-ISSN: 3025-9444

Comparison of Content-Based Filtering and K-Nearest Neighbor in Genre-Based Movie Recommendations

Irfan Nur Sofiyanto [1]✉, Nurmalitasari [2], Rudi Susanto [3]
(1,2,3) Faculty of Computer Science, Universitas Duta Bangsa, Surakarta, Indonesia

Abstract: *The rapid growth of digital movie platforms has increased the number of available movies, making it difficult for users to find films that match their preferences. This study aims to implement and compare Content-Based Filtering and K-Nearest Neighbor methods in a genre-based movie recommendation system. The data used in this study were obtained from the Full TMDB Movies Dataset from Kaggle, consisting of 1,422,047 initial records. The dataset was processed through several stages, including attribute selection, missing value handling, genre transformation, duplicate data removal, and feature weighting using TF-IDF, resulting in 796,425 movie records ready for modeling. Content-Based Filtering was implemented using cosine similarity to calculate the similarity between user genre preferences and movie features, while K-Nearest Neighbor was applied by identifying movies with the closest distance to the user preference vector. The results show that both methods were able to generate relevant movie recommendations based on the selected genres, namely Action, Adventure, and Science Fiction. Based on the evaluation results, both methods achieved the same accuracy of 99.92%. However, K-Nearest Neighbor showed slightly better performance, with a precision of 86.56%, recall of 80.32%, and F1-score of 83.32%, compared to Content-Based Filtering, which achieved a precision of 86.45%, recall of 79.58%, and F1-score of 82.87%. These findings indicate that K-Nearest Neighbor is slightly more effective in identifying relevant movie recommendations based on user genre preferences.*

Article history:
Received: 26 June 2026
Revised: 27 July 2026
Accepted: 29 July 2026
Published: 31 July 2026

Keyword:
content-based filtering, k-nearest neighbor, movie recommendation, machine learning

This is an open-access article under the CC-BY-SA License.

How to cite: Sofiyanto, I. N., Nurmalitasari, N., & Susanto, R. (2026). Comparison of Content-Based Filtering and K-Nearest Neighbor in Genre-Based Movie Recommendations. *RESWARA: Jurnal Riset Ilmu Teknik*, *4*(3), 1056–1069. https://doi.org/10.70716/reswara.v4i3.670

INTRODUCTION

The rapid development of information technology and the internet has significantly transformed the way people access entertainment, particularly movies. In the past, movies were primarily distributed through television, cinemas, and physical media such as DVDs. Today, digital platforms and streaming services enable users to access extensive movie collections anytime and anywhere (Riyadi et al., 2024). Although this technological advancement has improved accessibility, it has also introduced a new challenge. The vast number of available movies across different genres, countries, and production years often makes it difficult for users to identify movies that best match their interests and preferences (Aditi et al., 2024).

To address this challenge, recommendation systems have become an essential component of modern digital platforms. Recommendation systems assist users in discovering relevant content by analyzing available information, such as item characteristics, user interaction history, and similarity patterns among data (Nguyen et al., 2020; Bsoul et al., 2025). By reducing the need for users to manually browse extensive catalogs, recommendation systems improve user experience while increasing the efficiency of content discovery (Savitri et al., 2021). The recommendation process generally involves data collection, feature extraction, model development, and performance evaluation

✉ *Corresponding author:* ***220103179@mhs.udb.ac.id***
Copyright © 2026, the author(s)

to ensure that the generated recommendations accurately reflect user preferences (Andayani et al., 2024).

The implementation of recommendation systems is closely related to advances in machine learning, which enables computer systems to learn patterns from data and make predictions or recommendations with minimal human intervention (Nguyen et al., 2023). Various machine learning approaches, including supervised, unsupervised, and reinforcement learning, have been widely applied to recommendation tasks depending on the characteristics of the available data (Apmi Salma Kinanthi et al., 2025; Muliawan et al., 2022; Lonang & Normawati, 2022). Among these approaches, content-based recommendation techniques have gained considerable attention because they focus on analyzing item attributes to generate personalized recommendations.

One of the most widely used recommendation techniques is Content-Based Filtering (CBF), which recommends items based on the similarity of their attributes to those previously preferred by users (Cholik, 2021). By constructing user preference profiles from historical interactions, CBF can provide personalized recommendations for items with similar characteristics (Ardiansyah et al., 2023; Sitti Aliyah Azzahra et al., 2024). Another commonly applied approach is the K-Nearest Neighbor (KNN) algorithm, which determines recommendations by measuring the similarity or distance between data objects (Asa Dilla Safitri et al., 2024; Azizah & Rozi, 2024). Although both methods rely on similarity concepts, they employ different computational mechanisms, potentially resulting in different recommendation performance and relevance (Ridwansyah et al., 2024; Rochmad Wahono et al., 2024).

Previous studies have demonstrated the effectiveness of Content-Based Filtering and K-Nearest Neighbor in various recommendation domains. However, most existing studies evaluate these algorithms independently or focus on specific application scenarios, making direct comparisons under the same experimental conditions relatively limited. Furthermore, only a few studies comprehensively compare both approaches using identical datasets and genre-based movie features to determine their effectiveness in generating personalized movie recommendations. This limitation creates an opportunity to investigate the comparative performance of these algorithms using a consistent evaluation framework.

Therefore, this study aims to compare the performance of Content-Based Filtering and K-Nearest Neighbor in generating genre-based movie recommendations. The comparison is conducted using the same movie dataset and evaluation process to analyze the strengths and limitations of each algorithm in producing relevant recommendations according to user genre preferences. The findings of this study are expected to provide practical insights into selecting appropriate recommendation algorithms for movie recommendation systems and contribute to the development of more effective personalized recommendation models.

RESEARCH METHOD

This study employed a quantitative experimental approach to compare the performance of two recommendation algorithms, namely Content-Based Filtering (CBF) and K-Nearest Neighbor (KNN), in generating genre-based movie recommendations. Both algorithms were implemented using the same dataset, preprocessing procedures, feature extraction techniques, and evaluation metrics to ensure a fair and objective comparison. The overall research methodology consists of eight sequential stages, beginning with problem identification and ending with the research conclusion.

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

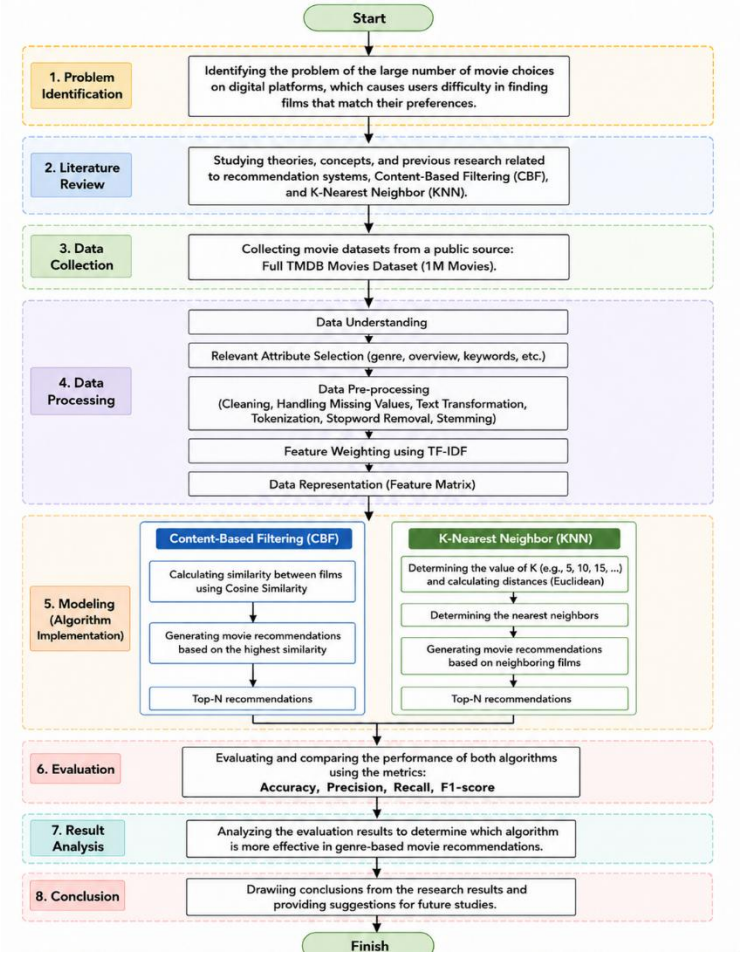

Figure 1. Research Framework

The research process begins with problem identification, where the increasing number of movies available on digital streaming platforms creates challenges for users in selecting movies that match their preferences. This problem motivates the development of an effective recommendation system.

The second stage is the literature review, which examines previous studies and theories related to recommendation systems, machine learning, Content-Based Filtering (CBF), and K-Nearest Neighbor (KNN). This stage provides the theoretical foundation for selecting the algorithms and identifying the research gap addressed in this study.

Next, the data collection stage gathers movie information from the Full TMDB Movies Dataset (1M Movies) obtained from Kaggle. The dataset contains comprehensive movie metadata, including movie titles, genres, overviews, keywords, ratings, and other descriptive attributes that are suitable for content-based recommendation.

The collected dataset is then processed through several data preprocessing stages. Initially, data understanding is conducted to examine the dataset structure and identify relevant information. Subsequently, attribute selection is performed to retain only the features required for recommendation generation, including movie titles, genres, overviews, and keywords. The textual data then undergo cleaning, missing value handling, lowercase conversion, tokenization, stop-word removal, and stemming or lemmatization where appropriate. Finally, the cleaned text is transformed

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) weighting scheme, producing a feature matrix that represents each movie.

The processed feature matrix is subsequently used during the model implementation stage. Two recommendation algorithms are independently implemented using identical feature representations. The Content-Based Filtering algorithm computes movie similarity using Cosine Similarity, ranks similarity scores, and generates Top-N movie recommendations based on the highest similarity values. Meanwhile, the K-Nearest Neighbor algorithm determines the optimal value of k, calculates the Euclidean distance between movie vectors, identifies the nearest neighboring movies, and produces Top-N recommendations based on the closest neighbors.

The recommendation results generated by both algorithms are evaluated during the performance evaluation stage using Accuracy, Precision, Recall, and F1-score. These evaluation metrics are used to objectively compare the effectiveness of the two recommendation approaches under identical experimental conditions.

Finally, the evaluation results are analyzed to determine which algorithm provides more relevant genre-based movie recommendations. Based on these findings, conclusions are drawn and recommendations for future research are proposed.

Research Design

This research adopted a quantitative experimental design to compare the effectiveness of Content-Based Filtering and K-Nearest Neighbor for genre-based movie recommendation. Both algorithms were evaluated under identical experimental settings using the same dataset, preprocessing procedures, feature extraction techniques, and evaluation metrics. This design ensures that the observed performance differences are attributable solely to the algorithms rather than differences in data preparation or evaluation methodology.

Dataset and Data Preprocessing

The dataset used in this study was the Full TMDB Movies Dataset (1M Movies) obtained from Kaggle. Only attributes relevant to recommendation generation, including movie title, genre, overview, and keywords, were selected for analysis.

The preprocessing stage consisted of data cleaning, duplicate removal, missing value handling, feature selection, lowercase conversion, tokenization, stop-word removal, and stemming or lemmatization where applicable. The processed textual data were transformed into numerical vectors using the TF-IDF weighting method:

$$\text{TFIDF}(t,d)=\text{TF}(t,d)\times\log\left(\frac{N}{DF(t)}\right)$$

where TF(t,d)TF(t,d)TF(t,d) represents the frequency of term ttt in document ddd, DF(t)DF(t)DF(t) denotes the number of documents containing term ttt, and NNN is the total number of documents.

Recommendation Algorithms

Two recommendation algorithms were implemented using the same TF-IDF feature representation.

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

The Content-Based Filtering (CBF) algorithm generates recommendations by calculating the similarity between movie feature vectors using Cosine Similarity:

$$\text{Cosine Similarity}(\mathrm{A}, \mathrm{B}) = \frac{\mathrm{A} \cdot \mathrm{B}}{\|\mathrm{A}\|\|\mathrm{B}\|}$$

where AAA and BBB denote the TF-IDF vectors of two movies. The movies with the highest similarity scores are ranked to produce the Top-N recommendations.

The K-Nearest Neighbor (KNN) algorithm recommends movies by identifying the nearest neighboring feature vectors. The Euclidean distance is calculated as:

$$d(x, y) = \backslash sqrt\left\{\backslash sum_{\{i=1\}}^{\{n\}(x_i - y_i)^2}\right\}$$

where $x_i - y_i$ and $x_i - y_i$ represent the feature values of two movie vectors. After determining the value of kkk, the nearest neighboring movies are selected to generate the recommendation list.

Performance Evaluation

The performance of both recommendation algorithms was evaluated using Accuracy, Precision, Recall, and F1-score. These metrics provide a comprehensive assessment of recommendation quality by measuring overall correctness, the proportion of relevant recommended movies, the ability to retrieve relevant movies, and the balance between precision and recall. The evaluation results were then compared to determine which algorithm provides more effective and relevant genre-based movie recommendations.

RESULTS AND DISCUSSION

Dataset Checking

This subsection explains the dataset used after collection, namely the Full TMDB Movies Dataset (1M Movies). This section explains the initial data set, available attributes, and the attributes used in the study.

Table 1. Dataset summary

Description	Value
Dataset Name	Full TMDB Movies Dataset (1M Movies)
Dataset Source	Kaggle
Data Format	CSV
Number of Records	1,422,047
Number of Attributes	24
Number of Missing Values	9,335,451
Duplicate Data Based on ID	1,245
Duplicate Data Based on Title	229,910

The available attributes in the dataset consist of 24 columns. Not all attributes are used directly in the recommendation process, as some are more administrative or supporting, such as homepage, poster_path, backdrop_path, and imdb_id. Meanwhile, the more relevant attributes for a movie recommendation system are title, genres, overview, keywords, popularity, vote_average, vote_count,

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

and release_date. These attributes can be used to represent film characteristics, both in terms of content and popularity.

Table 2. Numerical Attribute Statistics

Attribute	Number of Valid Data	Average	Minimum Value	Maximum Value
popularity	1,422,047	1.0094	0	2,994.36
vote_average	1,422,047	1.552	0	10
vote_count	1,422,047	15.0893	0	34,495
runtime	1,422,047	43.9142	-28	14,400
budget	1,422,047	235,561.84	0	1,000,000,000
revenue	1,422,047	620,210.81	-12	5,000,000,000

Based on the numerical attribute statistics, the popularity, vote_average, vote_count, runtime, budget, and revenue attributes have a total of 1,422,047 valid data points. The vote_average attribute has an average value of 1.5520, while the vote_count attribute has an average value of 15.0893. Several abnormal values were found in the runtime and revenue attributes, such as a minimum runtime value of -28 and a minimum revenue value of -12. These values indicate the possibility of recording errors in the dataset. However, runtime, budget, and revenue were not used as main features in the recommendation modeling process because this study focuses on genre-based movie recommendations. Therefore, the data preparation stage focused on attributes directly related to the recommendation process, particularly title and genre, while numerical attributes such as popularity, vote_average, and vote_count were used only as supporting information in displaying the recommendation results.

Table 3. Film Genre Distribution

Genre	Number of Movies
Drama	261,365
Documentary	197,705
Comedy	159,332
Animation	67,296
Horror	64,679
Romance	60,152
Music	56,962
Thriller	55,058
Action	50,156
Crime	37,895
Family	31,403
Fantasy	27,624
Adventure	27,025
TV Movie	25,902
Science Fiction	25,465
Mystery	24,164
History	18,973
War	11,674
Western	9,493

The genre distribution shows that the dataset has quite a diverse range of genres. The genre with the largest number of films is Drama, with 261,365 films, followed by Documentary with 197,705 films, and Comedy with 159,332 films. The dominance of Drama, Documentary, and Comedy genres

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

indicates that the dataset has a large representation of common genres. Meanwhile, genres such as Western, War, and History have fewer data sets than other genres.

Data Preprocessing

The data preparation stage is carried out to prepare the data for use in the recommendation system modeling process. The dataset used is still in raw form, requiring cleaning and format adjustments. This stage aims to make the data more structured, relevant, and ready to be processed by the Content-Based Filtering and K-Nearest Neighbor algorithms.

Table 4. Data Amount After Data Preprocessing

Stage	Number of Data	Description
Initial Data	1,422,047	Total number of data in the initial dataset
Attribute Selection	1,422,047	Data after relevant attributes were selected
Data Removed Due to Missing Title	25	Data without movie titles
Data Removed Due to Missing Genre	625,481	Data without genre information
Duplicate Data Removed	116	Data with the same ID or the same combination of title and year
Final Prepared Data	796,425	Data ready to be used for modeling

Based on the data preparation results, the dataset has gone through several important stages: attribute selection, empty data cleaning, genre transformation, duplicate data removal, recommendation feature formation, and feature weighting using TF-IDF. After these processes, the final dataset ready for use was 796,425 films. Data without titles and without genres were removed, as these two attributes play a crucial role in the recommendation process. Furthermore, duplicate data was also removed to avoid data redundancy, which could impact recommendation results.

The genre attribute, which previously had a diverse format, was converted to a cleaner and more uniform format using the genre_clean attribute. This attribute was then used to form feature_text, the primary feature in genre-based recommendations. The text features were then transformed into numeric form using TF-IDF so they could be processed by the algorithm. Thus, the data used was clean, structured, free of gaps in the primary attributes, and represented in a feature matrix. Therefore, the dataset is ready to enter the modeling stage, namely the application of the Content-Based Filtering and K-Nearest Neighbor algorithms to generate movie recommendations.

Feature Weighting Using TF-IDF

After data preprocessing was carried out, the next stage was feature weighting using TF-IDF. The following table presents an example of the TF-IDF representation results for several movie data, where each row represents one movie and each column represents the genre features used in the modeling process. The value in each column indicates the weight level of a particular genre for a specific movie.

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

Tabel 5. Feature Wighting Using TF-IDF

Index Film	action	adventure	comedy	crime	drama	fantasy	science_fiction	thriller
0	0,517	0,601	0	0	0	0	0,61	0
1	0	0,665	0	0	0,321	0	0,674	0
2	0,541	0	0	0,581	0,304	0	0	0,527
3	0,443	0,516	0	0	0	0,513	0,523	0
4	0,517	0,601	0	0	0	0	0,61	0
5	0,594	0,691	0,412	0	0	0	0	0
6	0,517	0,601	0	0	0	0	0,61	0
7	0	0	0	0	1	0	0	0
8	0,517	0,601	0	0	0	0	0,61	0
9	0	0	0	0,74	0	0	0	0,672

The TF-IDF data representation table shows that each movie has been transformed into a numerical vector based on its genre features. Each row represents a movie, while each column represents a genre. A value of 0 indicates that the genre is not present in the movie, while a value above 0 indicates that the genre contributes to the movie representation.

Based on the displayed data, several movies have similar genre patterns, such as movies with indexes 0, 4, 6, and 8, which share the genres action, adventure, and science fiction. This similarity indicates that these movies have similar characteristics and may be recommended to one another. Meanwhile, the movie with index 7 only has a weight in the drama genre, meaning it is represented as a single-genre movie. The TF-IDF representation is then used as the basis for calculating movie similarity. In Content-Based Filtering, the TF-IDF vectors are used with cosine similarity, while in K-Nearest Neighbor, they are used to identify the closest movies based on genre feature similarity.

Implementation Content Based Filtering Algorithm

Content-Based Filtering is a recommendation system method used to generate recommendations based on the similarity of characteristics between items. In this study, the Content-Based Filtering method is applied to recommend movies based on genre similarity, which has been represented in the form of a TF-IDF matrix. Each movie is represented as a numerical vector and then compared with other movies to determine the level of similarity based on its genre features. The similarity calculation process is carried out using Cosine Similarity, so movies with the highest similarity scores are selected as the recommendation results.

	id	Title	Genres	Language	Release Date	Release Year	Vote Average	Vote Count	Popularity	Genre Match Count	Genre Match Ratio	CBF Score
0	1124195	Doctor Who: The Christmas Invasion	Science Fiction, Adventure, Action	en	2005-12-25	2005.0	10.0	1	2.355	3	1.0	1.0
1	509300	The Legend of Space Sheriff Gavan	Action, Adventure, Science Fiction	ja	2012-10-20	2012.0	10.0	1	1.222	3	1.0	1.0
2	584605	Kamen Rider Kuuga: Special Edition	Action, Adventure, Science Fiction	ja	2001-10-21	2001.0	10.0	1	1.062	3	1.0	1.0
3	587597	Ultra Fight Victory	Action, Adventure, Science Fiction	ja	2016-07-05	2016.0	10.0	1	0.988	3	1.0	1.0
4	433560	Himitsu Sentai Gorenger: The Blue Fortress	Science Fiction, Action, Adventure	ja	1975-07-26	1975.0	10.0	1	0.879	3	1.0	1.0
5	994830	Coal Miner's Adventure	Action, Adventure, Science Fiction	en	NaT	NaN	10.0	1	0.870	3	1.0	1.0
6	433564	Himitsu Sentai Gorenger: The Red Death Match	Science Fiction, Action, Adventure	ja	1976-01-10	1976.0	10.0	1	0.677	3	1.0	1.0
7	217928	XX Ray II	Action, Science Fiction, Adventure	ms	1995-11-24	1995.0	10.0	1	0.623	3	1.0	1.0
8	647225	The Traveler	Action, Adventure, Science Fiction	en	2019-11-12	2019.0	10.0	1	0.600	3	1.0	1.0
9	521320	Blood and Steel	Science Fiction, Action, Adventure	en	2015-06-01	2015.0	10.0	1	0.600	3	1.0	1.0

Figure 2. Result Content Based Filtering Algorithm

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

Based on the Content-Based Filtering recommendation results, the system generated a list of movies that have genre similarities with the user's preferences, namely Action, Adventure, and Science Fiction. All movies in the recommendation list have a Genre Match Count value of 3 and a Genre Match Ratio of 1.0, indicating that all user-selected genres are present in the recommended movies. In addition, the CBF Score of 1.0 shows the maximum level of similarity between the user's preferences and the recommended movies. These results indicate that the Content-Based Filtering method is able to provide relevant movie recommendations based on genre similarity. However, most of the movies have a low number of votes, so the evaluation of movie quality based on rating needs to be analyzed carefully because a high vote_average does not necessarily reflect broad popularity or user acceptance.

Implementation KNN Algorithm

K-Nearest Neighbor (KNN) is an algorithm used to determine recommendations based on data proximity or similarity. In this study, the value of K in the K-Nearest Neighbor method was determined based on the number of nearest movies selected as recommendation results. Several K values were tested to observe the consistency of recommendation relevance, including K = 5, K = 10, K = 15, and K = 20. The evaluation was conducted by observing the genre match ratio and the distance value between the user preference vector and the movie vectors. Based on the testing results, K = 10 was used in this study because it provides a sufficient number of recommendations while still maintaining high genre relevance. Therefore, the KNN algorithm generates the top 10 nearest movies based on the smallest distance values.

	id	Title	Genres	Language	Release Date	Release Year	Vote Average	Vote Count	Popularity	Genre Match Count	Genre Match Ratio	KNN Distance	KNN Score
0	1124195	Doctor Who: The Christmas Invasion	Science Fiction, Adventure, Action	en	2005-12-25	2005.0	10.0	1	2.355	3	1.0	0.0	1.0
1	509300	The Legend of Space Sheriff Gavan	Action, Adventure, Science Fiction	ja	2012-10-20	2012.0	10.0	1	1.222	3	1.0	0.0	1.0
2	584605	Kamen Rider Kuuga: Special Edition	Action, Adventure, Science Fiction	ja	2001-10-21	2001.0	10.0	1	1.062	3	1.0	0.0	1.0
3	587597	Ultra Fight Victory	Action, Adventure, Science Fiction	ja	2016-07-05	2016.0	10.0	1	0.988	3	1.0	0.0	1.0
4	433560	Himitsu Sentai Gorenger: The Blue Fortress	Science Fiction, Action, Adventure	ja	1975-07-26	1975.0	10.0	1	0.879	3	1.0	0.0	1.0
5	994830	Coal Miner's Adventure	Action, Adventure, Science Fiction	en	NaT	NaN	10.0	1	0.870	3	1.0	0.0	1.0
6	433564	Himitsu Sentai Gorenger: The Red Death Match	Science Fiction, Action, Adventure	ja	1976-01-10	1976.0	10.0	1	0.677	3	1.0	0.0	1.0
7	217928	XX Ray II	Action, Science Fiction, Adventure	ms	1995-11-24	1995.0	10.0	1	0.623	3	1.0	0.0	1.0
8	647225	The Traveler	Action, Adventure, Science Fiction	en	2019-11-12	2019.0	10.0	1	0.600	3	1.0	0.0	1.0
9	521320	Blood and Steel	Science Fiction, Action, Adventure	en	2015-06-01	2015.0	10.0	1	0.600	3	1.0	0.0	1.0

Figure 3. Result KNN Algorithm

Based on the K-Nearest Neighbor recommendation results, the system generated a list of movies with the highest proximity to the user's genre preferences, namely Action, Adventure, and Science Fiction. All movies in the recommendation list have a Genre Match Count value of 3 and a Genre Match Ratio of 1.0, indicating that all recommended movies fully match the selected genres. The KNN Distance value of 0.0 indicates that the distance between the user preference vector and the movie vector is zero, meaning that these movies have the same genre representation as the user's preferences. In addition, the KNN Score of 1.0 indicates the maximum level of proximity. These results show that the K-Nearest Neighbor method is able to generate relevant recommendations based on the proximity of genre features, although movie ratings still need to be analyzed carefully because most of the movies have a low number of votes.

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

Comparison Ranking of Content-Basef Filtering and KNN

After the recommendation results from the Content-Based Filtering and K-Nearest Neighbor methods were obtained separately, the next stage was to compare the recommendation rankings of both methods. This comparison was conducted to determine whether the two algorithms produced the same or different movie recommendation orders based on the user's genre preferences. At this stage, the recommendation results from each method were compared based on ranking position, recommended movie title, CBF score, KNN distance, and KNN score. The results of this comparison were used to examine the similarity level of the recommendation outputs before further model performance evaluation was carried out.

	Rank	CBF Title	CBF Score	KNN Title	KNN Distance	KNN Score	Status
0	1	Doctor Who: The Christmas Invasion	1.0	Doctor Who: The Christmas Invasion	0.0	1.0	Same
1	2	The Legend of Space Sheriff Gavan	1.0	The Legend of Space Sheriff Gavan	0.0	1.0	Same
2	3	Kamen Rider Kuuga: Special Edition	1.0	Kamen Rider Kuuga: Special Edition	0.0	1.0	Same
3	4	Ultra Fight Victory	1.0	Ultra Fight Victory	0.0	1.0	Same
4	5	Himitsu Sentai Gorenger: The Blue Fortress	1.0	Himitsu Sentai Gorenger: The Blue Fortress	0.0	1.0	Same
5	6	Coal Miner's Adventure	1.0	Coal Miner's Adventure	0.0	1.0	Same
6	7	Himitsu Sentai Gorenger: The Red Death Match	1.0	Himitsu Sentai Gorenger: The Red Death Match	0.0	1.0	Same
7	8	XX Ray II	1.0	XX Ray II	0.0	1.0	Same
8	9	The Traveler	1.0	The Traveler	0.0	1.0	Same
9	10	Blood and Steel	1.0	Blood and Steel	0.0	1.0	Same

Figure 4. Comparison Ranking of Content-Based Filtering and KNN

Confusion Matrix and Evaluation Metric Comparison

The confusion matrix is used to determine the model's ability to classify recommendation data into relevant and irrelevant categories based on the user's genre preferences. In this study, the confusion matrix was applied to the Content-Based Filtering and K-Nearest Neighbor methods to identify the number of correct and incorrect predictions through True Positive, True Negative, False Positive, and False Negative values. The results of the confusion matrix were then used as the basis for comparing evaluation metrics, namely accuracy, precision, recall, and F1-score. This comparison aims to determine which method has better performance in generating relevant movie recommendations based on the suitability of the user's genre preferences.

Table 6. Confusion Matrix CBF and KNN

CONFUSION MATRIX CBF		
Actual / Predicted	Predicted Irrelevant	Predicted Relevant
Actual Irrelevant	794288	237
Actual Relevant	388	1512
CONFUSION MATRIX KNN		
Actual / Predicted	Predicted Irrelevant	Predicted Relevant
Actual Irrelevant	794288	237
Actual Relevant	374	1526

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

Based on the confusion matrix results, the Content-Based Filtering method produced 794,288 True Negatives, 237 False Positives, 388 False Negatives, and 1,512 True Positives. These results indicate that most irrelevant data were correctly classified as irrelevant, while some relevant data were still not accurately detected. Meanwhile, the K-Nearest Neighbor method produced 794,288 True Negatives, 237 False Positives, 374 False Negatives, and 1,526 True Positives. In comparison, KNN has a higher number of True Positives and a lower number of False Negatives than Content-Based Filtering. This indicates that KNN is slightly better at identifying relevant data based on the user's genre preferences.

Table 7. Comparison Metrix Evaluation

No	Model	Threshold	True Negative	False Positive	False Negative	True Positive	Accuracy	Accuracy (%)	Precision	Precision (%)	Recall	Recall (%)	F1-Score	F1-Score (%)
0	Content-Based Filtering	0.82	794288	237	388	1512	0.9992	99.92	0.8645	86.45	0.7958	79.58	0.8287	82.87
1	K-Nearest Neighbor	0.62	794288	237	374	1526	0.9992	99.92	0.8656	86.56	0.8032	80.32	0.8332	83.32

Based on the evaluation metric comparison results, both the Content-Based Filtering and K-Nearest Neighbor methods achieved an accuracy value of 99.92%. This indicates that both methods have a very good ability to classify relevant and irrelevant data. However, when viewed from the precision, recall, and F1-score metrics, the K-Nearest Neighbor method shows slightly better results. KNN achieved a precision of 86.56%, a recall of 80.32%, and an F1-score of 83.32%, while Content-Based Filtering achieved a precision of 86.45%, a recall of 79.58%, and an F1-score of 82.87%. In addition, KNN produced a higher True Positive value and a lower False Negative value than CBF. These results indicate that KNN is slightly superior in identifying relevant movies based on the user's genre preferences.

Tabel 8. Best Model Comparison

No	Aspek	Model Terbaik
0	Accuracy	Seimbang
1	Precision	K-Nearest Neighbor
2	Recall	K-Nearest Neighbor
3	F1-Score	K-Nearest Neighbor
4	Waktu Proses	K-Nearest Neighbor

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

Based on the best model based on metrics table, it can be seen that both methods have balanced performance in terms of accuracy. This indicates that Content-Based Filtering and K-Nearest Neighbor have equally good abilities in classifying recommendation data in general.

However, in terms of precision, recall, F1-score, and processing time, the K-Nearest Neighbor method is identified as the best model. The higher precision value indicates that KNN is more accurate in generating movie recommendations that are truly relevant. The better recall value shows that KNN is more capable of identifying relevant movies according to the user's genre preferences. Meanwhile, the higher F1-score indicates that KNN has a better balance between precision and recall.

In addition, KNN also performs better in terms of processing time, meaning that this method is more efficient in generating recommendations compared to the other method. Based on these results, K-Nearest Neighbor can be considered the better method in this study because it outperforms the other method in most evaluation metrics.

CONCLUSION

Based on the research results, it can be concluded that a genre-based movie recommendation system was successfully developed using Content-Based Filtering and K-Nearest Neighbor algorithms. The dataset used in this study was obtained from the Full TMDB Movies Dataset and went through several data preparation stages, including attribute selection, missing value handling, genre transformation, duplicate data removal, and feature weighting using TF-IDF. After the data preparation process, 796,425 movie records were obtained and used in the modeling stage.

The results show that both Content-Based Filtering and K-Nearest Neighbor were able to generate movie recommendations based on user genre preferences. In the recommendation results, both methods produced movies that matched the selected genres, namely Action, Adventure, and Science Fiction. Content-Based Filtering generated recommendations based on cosine similarity between the user preference vector and the movie feature vectors, while K-Nearest Neighbor generated recommendations based on the closest distance between the user preference vector and movie vectors.

Based on the confusion matrix and evaluation metrics, both methods achieved the same accuracy value of 99.92%, indicating that both algorithms have a very good ability to classify relevant and irrelevant movie data. However, K-Nearest Neighbor showed slightly better performance in precision, recall, and F1-score. KNN achieved a precision of 86.56%, recall of 80.32%, and F1-score of 83.32%, while Content-Based Filtering achieved a precision of 86.45%, recall of 79.58%, and F1-score of 82.87%. In addition, KNN produced a higher True Positive value and a lower False Negative value than Content-Based Filtering.

These results indicate that K-Nearest Neighbor is slightly more effective than Content-Based Filtering in identifying relevant movies based on user genre preferences. Although the difference between the two methods is relatively small, KNN can be considered the better-performing method in this study because it outperforms Content-Based Filtering in most evaluation metrics, especially precision, recall, F1-score, and the detection of relevant movie recommendations.

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

REFERENCES

Aditi, A., Dubey, G. P., Shakya, H. K., & Sharma, A. (2024). SOM and Hybrid Filtering: Pioneering Next-Gen Movie Recommendations in the Entertainment Industry. *Fusion: Practice and Applications*, *16*(2), 43–62. https://doi.org/10.54216/FPA.160204

Andayani, D. D., Nurfitri, A. A., Adiba, F., & Yahya, M. (2024). *Early detection system for melanoma skin cancer using reinforcement learning method*. 030005. https://doi.org/10.1063/5.0221127

Apmi Salma Kinanthi, Alivia Putrian Wijaya, Dwi Yanawati, Siti Syamsidar, & Edy Widodo. (2025). Pengelompokkan Kabupaten/Kota di Pulau Jawa Berdasarkan Faktor Kemiskinan Menggunakan Metode Hierarchical Clustering. *Evolusi : Jurnal Sains Dan Manajemen*, *13*(1), 40–51. https://doi.org/10.31294/evolusi.v13i1.8209

Ardiansyah, R., Ari Bianto, M., & Saputra, B. D. (2023). Sistem Rekomendasi Buku Perpustakaan Sekolah menggunakan Metode Content-Based Filtering. *Jurnal CoSciTech (Computer Science and Information Technology)*, *4*(2), 510–518. https://doi.org/10.37859/coscitech.v4i2.5131

Asa Dilla Safitri, Vihi Atina, & Anisatul Farida. (2024). Sistem rekomendasi buku menggunakan metode content-based filtering. *INFOTECH : Jurnal Informatika & Teknologi*, *5*(2), 218–227. https://doi.org/10.37373/infotech.v5i2.1302

Azizah, N., & Rozi, A. F. (2024). Sistem Rekomendasi Produk Somethinc Menggunakan Metode Content-based Filtering. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, *6*(3), 461–468. https://doi.org/10.47233/jteksis.v6i3.1411

Bsoul, Q., Zawaideh, F., Alqadi, B. S., Almusfar, L. A., Khalaf, O. I., Alattas, A. S., Alali, M., & AbdElminaam, D. S. (2025). From User Preferences to Accurate Predictions: Enhancing Movie Recommendation Systems with Neural Collaborative Filtering and Sentiment Analysis. *SN Computer Science*, *6*(3), 257. https://doi.org/10.1007/s42979-025-03742-7

Cholik, C. A. (2021). Perkembangan Teknologi Informasi Komunikasi / ICT dalam Berbagai Bidang. *Jurnal Fakultas Teknik UNISA Kuningan*, *2*(2), 39–46.

Huda, A. A., Fajarudin, R., & Hadinegoro, A. (2022). Sistem Rekomendasi Content-based Filtering Menggunakan TF-IDF Vector Similarity Untuk Rekomendasi Artikel Berita. *Building of Informatics, Technology and Science (BITS)*, *4*(3). https://doi.org/10.47065/bits.v4i3.2511

Kusuma, I. H., & Cahyono, N. (2023). Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor. *Jurnal Informatika: Jurnal Pengembangan IT*, *8*(3), 302–307. https://doi.org/10.30591/jpit.v8i3.5734

Lonang, S., & Normawati, D. (2022). Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(1), 49. https://doi.org/10.30865/mib.v6i1.3312

Muliawan, A., Badriyah, T., & Syarif, I. (2022). Membangun Sistem Rekomendasi Hotel dengan Content Based Filtering Menggunakan K-Nearest Neighbor dan Haversine Formula. *Technomedia Journal*, *7*(2), 231–247. https://doi.org/10.33050/tmj.v7i2.1893

Nguyen, L. V., Nguyen, T.-H., & Jung, J. J. (2020). Content-Based Collaborative Filtering using Word Embedding. *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, 96–100. https://doi.org/10.1145/3400286.3418253

Nguyen, L. V., Vo, Q.-T., & Nguyen, T.-H. (2023). Adaptive KNN-Based Extended Collaborative Filtering Recommendation Services. *Big Data and Cognitive Computing*, *7*(2), 106. https://doi.org/10.3390/bdcc7020106

Retnoningsih, E., & Pramudita, R. (2020). Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python. *BINA INSANI ICT JOURNAL*, *7*(2), 156. https://doi.org/10.51211/biict.v7i2.1422

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.

Ridwansyah, T., Subartini, B., & Sylviani, S. (2024). Penerapan Metode Content-Based Filtering pada Sistem Rekomendasi. *Mathematical Sciences and Applications Journal*, *4*(2), 70–77. https://doi.org/10.22437/msa.v4i2.32136

Riyadi, S., Audita, N., & Abidin, F. A. (2024). *Movie Recommendation Using Collaborative Filtering Method and K-Nearest Neighbours: A Case Study on Netflix* (pp. 428–436). https://doi.org/10.1007/978-3-031-70518-2_38

Rochmad Wahono, A., Aji Saputra, B., & Fadlu Rahman, F. (2024). Sistem Rekomendasi Film Menggunakan Metode Content-Based Filtering dan Algoritma K-Nearest Neighbors (KNN). *Prosiding Seminar Nasional Teknologi Informasi Dan Bisnis*, 1–6. https://doi.org/10.47701/senatib.v4i1.3994

Sari, N. noor kamala, Licantik, L., & Zahra, M. (2024). Pemanfaatan Sistem Rekomendasi Menggunakan Content-Based Filtering pada Hotel di Palangka Raya. *Technologia : Jurnal Ilmiah*, *15*(4), 754. https://doi.org/10.31602/tji.v15i4.16282

Savitri, N. L. P. C., Rahman, R. A., Venyutzky, R., & Rakhmawati, N. A. (2021). Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning. *Jurnal Teknik Informatika Dan Sistem Informasi*, *7*(1). https://doi.org/10.28932/jutisi.v7i1.3216

Sitti Aliyah Azzahra, Syafran Nurrahman, & Aep Saefullah. (2024). Integrasi Kecerdasan Buatan Dalam Sistem Rekomendasi Produk Untuk E-Commerce. *Jurnal Sains Dan Teknologi*, *3*(1), 21–28. https://doi.org/10.58169/saintek.v3i1.394

Copyright © 2026, the author(s). This is an open access article under the CC-BY-SA License.