Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 1 Iss. 1 February 2024, pp: 16-22
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Performance Comparison K-Nearest Neighbor, Naive Bayes, and Decision Tree Algorithms for Netflix Rating Classification

Zulkarnain[1*], Risma Mutia[2], Jane Astrid Ariani[3], Zidny Alfian Barik[4], Habil azmi[5]

[1,2,3]Department of Information Systems, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[4]Faculty of Ushuluddin, Al-Azhar University, Egypt
[5]Department of Comparative Mazhab, Faculty of Syari'ah Wal Qonun, Al-Azhar University, Egypt

E-Mail: [1]12150314261@students.uin-suska.ac.id,
[2]12150320149@students.uin-suska.ac.id, [3]12150320044@students.uin-suska.ac.id,
[4]zidnyalfianbarik123@gmail.com, [5]habilazmi02@gmail.com

Received Dec 04th 2023; Revised Dec 15th 2023; Accepted Jan 04th 2024
Corresponding Author: Zulkarnain

Abstract

Netflix is a streaming service platform that is growing along with the increasing number of internet users. This research aims to classify movie and TV show rating datasets on Netflix by comparing the KNN, Naive Bayes and Decision Tree algorithms to determine the accuracy comparison of the three algorithms. From the results of the analysis, it is found that the three algorithms produce a comparison of the accuracy of movie and tv show rating classification data on Netflix with different values. Based on the confusion matrix, namely Accuracy, Precision, and Recall, it is found that the Naive Bayes algorithm has the highest accuracy of 72%, the Decision Tree algorithm is 70% and the KNN algorithm has the lowest accuracy of 61%. From these results it can be stated that the Naive Bayes algorithm can classify movie and tv show rating data on Netflix better than compared to the other two algorithms.

Keyword: Classification, Decision Tree, K-Nearest Neighbor, Naïve Bayes, Netflix

1. INTRODUCTION

The increase in the number of Internet users has had a positive effect on the birth of various innovations by service providers, such as electronic shopping, online news portals and video streaming services or video-on-demand platforms such as Netflix [1]. Founded as a DVD rental company in the United States in the 1990s, Netflix is now a computer platform that allows nearly 130 million users in 190 countries to enjoy hundreds of thousands of hours of series and films [2]. The success of this service depends not only on the variety of its content but also on its ability to present ideal recommendations to users according to their preferences. The development of the Internet also affects data science, one of which is data mining.

Data science consists of a combination of classic scientific disciplines such as databases, statistics, distributed systems and data mining [3]. Data science includes a set of basic principles that support and complement the principles of deriving news and information from data [4]. The term most closely related to data science is data mining. Data mining is the process of gathering information using existing technology to obtain new information [5]. There are hundreds of different data mining algorithms and many details in their implementation. This analysis process uses three data mining algorithms to classify Netflix movie rating data.

Some of the data mining algorithms that are always used for data classification are K-Nearest Neighbors (KNN), Naive Bayes and Decision Tree [6]. The KNN algorithm (KNN) is one of the most popular algorithms due to its comprehensive features, simplicity and accuracy [7]. KNN is widely used in big data research and implementation due to its efficiency in classification, regression and clustering tasks [8]. Meanwhile, the Naive Bayes Algorithm is, as the name suggests, derived from the words Baye and theorem, which is known for the simplicity, efficiency and reliability of its algorithm [9]. Calculating posterior probabilities based on previous probabilities is made possible by Bayes and the theorem. Because it is assumed that the value of one feature does not depend on the value of other characteristics, the calculation of this model becomes simple [10]. A decision tree algorithm's structure is comparable to a flowchart; each node indicates the testing of a particular attribute (question), each test result generates a new branch (tree level), and each leaf node symbolizes a token

for that class. Portrayal [11]. Decision tree algorithms can more quickly classify data from large data sets and handle independent data functions [12].

All three mentioned algorithms have different characteristics. So this study compares the three, that is, compares the level of accuracy to determine which classification method is the best and optimal [13].

This study uses Netflix movie and TV rating data with xlsx data type, which has 11 attributes in the data. This information covers various movies and TV shows published and broadcast on Netflix between 2020 and 2022. The attribute imdb-score is an attribute marked as a result of a survey of bad and fair ratings for a movie or TV show. or a good degree where every attribute value is represented in the data.

In this research, the researcher's main objective is to explore and compare the performance of three data mining algorithms, namely KNN, Naive Bayes, and Decision Tree, in classifying film and TV show rating data on the Netflix platform. By involving rating data covering the period from 2020 to 2022, this research aims to provide in-depth insight into the ability of each algorithm to predict rating categories including poor, fair, or good. By better understanding the advantages and disadvantages of each algorithm, it is hoped that this research can contribute to the development of more optimal classification methods for rating data on streaming service platforms such as Netflix. In other words, the focus of this research is to provide a clearer picture of the effectiveness and consistency of each algorithm in the context of film and TV show rating classification, guiding the selection of the most appropriate methods for further analysis.

2. MATERIAL AND METHOD

In this research, the first stage is data collection, the dataset applied is movie and TV show data on the Netflix platform downloaded from the Kaggle website. Next, the data preprocessing process is carried out by normalizing the data using the min-max method. Then the data classification stage is carried out using 3 algorithms, namely KNN, Naive Bayes and Decision Tree, so as to get the accuracy results of the 3 algorithms. Then the accuracy results will be analyzed and concluded. This research methodology can be seen in Figure 1.

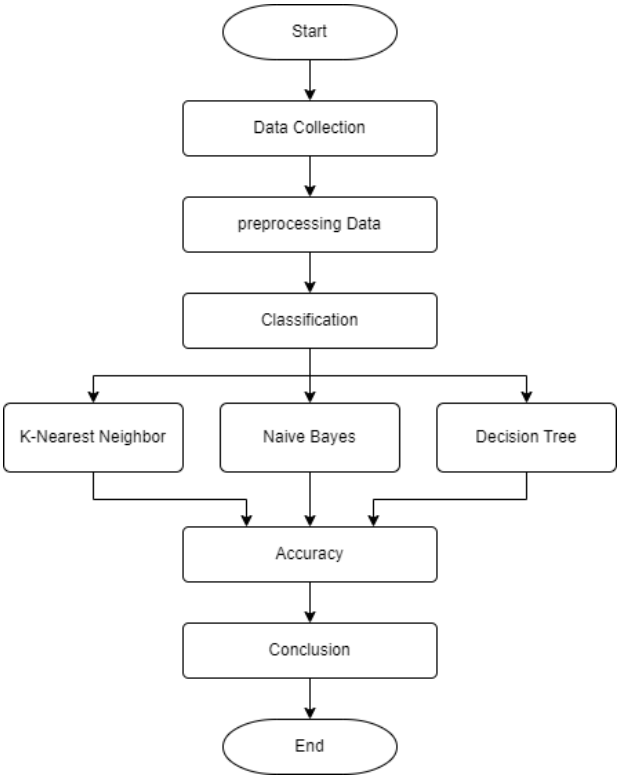

Figure 1. Research Methodology

2.1. Data Collecting

This research uses a dataset downloaded from Kaggle.com, which presents several attributes that include critical news for analyzing the classification of movies and TV shows on the Netflix platform. Index, ID, title, type, release year, duration, production country genre, IMDB ID, score and IMDb votes are used. Before being analyzed, this dataset has gone through a number of data preprocessing processes. This includes handling missing values on certain attributes, normalizing data for scale consistency, and converting categorical

attributes such as type and genre into numerical representations to account for and support classification algorithms [14].

2.2. K-Nearest Neighbor (KNN)

The KNN learning method is known as one of the simplest and easiest to understand algorithms [15]. In classification tasks, KNN classifies unlabeled queries or test samples by taking most of the k-nearest neighbors from different classes (using a uniform selection model). In this process, the nearest neighbor is selected based on a distance matrix or degree of dissimilarity, which depends on the type of attribute in question [16]. In the KNN classifier, the distance between samples is calculated using the Euclidean formula 1.

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (1)$$

Description:

d	: Distance
p	: Data Dimension
i	: Data Variable
x1	: Data Sample
x2	: Test Data

2.3. Naive Bayes

Calculating probabilities by calculating frequencies and combining values from a given data set is how the Naive Bayes probability classification algorithm is used [16]. In performing classifiers, this algorithm adopts the concept of mixture models, mixture models are able to determine probabilities from components consisting of the application of Bayes' theorem to serve as probability-based classifiers [17]. The following is the Naive Bayes formula 2.

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad (2)$$

Description:

X	: Data that has an unknown class
H	: Hypothesized data X is a specific class
P(H | X)	: Hypothesis H has probability based on conditions
P(X | H)	: probability X which is based on the conditions that exist in hypothesis H
P(H)	: Probability of hypothesis H (prior probability)
P(X)	: Chance of X

2.4. Decision Tree

Decision Tree, as the name suggests, is a classification method that is arranged in the form of a structure that resembles a tree [18]. Decision Trees are created from Root Node to Leaf Node through a recursive process. Each branch in the tree reflects a condition that must be met, while the tip of the tree shows the value of the relevant data [19]. The formula for calculating sample entropy on a decision tree is:

$$\text{Entropy}(S) = -P_1 \log_2 P1 - P2 \log_2 P2 \quad (1)$$

Where p1, p2, p3, ..., pn indicate propositions in class 1, class 2, ..., class n of the output, respectively

2.5. Google Collab

Google Colab, often called Google Colaboratory, is an open source service provided by Google to those who have a Gmail account [20]. Google Colab is increasingly used in educational and training environments due to its focus on knowledge dissemination and research in the field of machine learning [21]. With Google Collab, developers can create, modify and run code using Python programming languages such as NumPy and Matplotlib for data analysis and visualization [22].

3. RESULTS AND DISCUSSION

3.1. Data Collection

This study uses imdb TV SHOW and MOVIE rating data on the Netflix platform downloaded from the kaggle website. This information shows that the attributes used to rate TV shows and movies on the Netflix platform are index, identifier, name, type, production countries, release year, runtime, genres, imdb_id, imdb_score, imdb_votes. The determination of the accuracy value in this study relies on attribute classification using IMDb rating data for TV shows and movies available on the Netflix platform.

Table 1. Data Collection

Index	Id	Title	Type	Release Year	Runtime	Genres	Production Countries	Imdb id	Imdb Votes	Imdb Score
1	1	1	1	1	1	1	1	1	1	7,3
2	2	2	2	2	2	2	2	2	2	5,7
3	3	3	2	1	3	3	3	3	3	6,7
4	4	4	2	2	4	4	4	4	4	3,3
5	5	5	2	2	5	5	5	5	5	8,6
…	…	…	…	…	…	…	…	…	…	…
1427	1427	1427	2	1	105	525	9	1427	1256	5,5

3.2. Processing Data

The next stage is data preprocessing. Preprocessing is done by removing noise data such as data invalidation, blank data, typing errors, and so on. At this stage it will be ensured that the existing data records are not listed in the noise category. The amount of data used in this study amounted to 1427 data records. From the whole data, the transformation process will then be carried out and also the normalization of data which will later be used in the process using Google Collab. This stage transforms the imdb score attribute into 3 categories, where the "Excellent" category data totals 406, the "Good" category data totals 968 and the "Bad" category data totals 53 on the imdb score attribute.

Table 2. Preprocessing Data

Index	Id	Title	Type	Release Year	Runtime	Genres	Production Countries	Imdb id	Imdb Votes	Imdb Score
1	1	1	1	1	1	1	1	1	1	Excellent
2	2	2	2	2	2	2	2	2	2	Good
3	3	3	2	1	3	3	3	3	3	Good
4	4	4	2	2	4	4	4	4	4	Bad
5	5	5	2	2	5	5	5	5	5	Excellent
…	…	…	…	…	…	…	…	…	…	…
1427	1427	1427	2	1	105	525	9	1427	1256	Excellent

3.3. K-Nearest Neighbor (KNN)

In this study, the classification process is carried out using the K-nearest neighbor algorithm. Following are the results of KNN algorithm using imdb TV SHOW and MOVIE rating data on Netflix platform in Table 3.

Table 3. Result from KNN Performance Model

	True Good	True Excellent	True Bad	Class Precision
Pred. Good	156	47	1	69%
Pred. Excellent	58	12	0	20%
Pred. Bad	11	1	0	0%
Class Recall	76%	17%	0%	

From the table of Test results using the KNN algorithm above, it can be seen that the performance varies in identifying different classes. For the "Good" class, the model has a good precision of 69% and recall of 76%, indicating that most of the positive predictions are correct. However, the low precision for the "Excellent" class of 17% and the low recall values for the "Excellent" class of 20% and "bad" class of 0% indicate that the model tends to struggle in recognizing and distinguishing these classes.

3.4. Naive Bayes

The next classification process is to use the Naive Bayes algorithm. The results of testing with this algorithm can be seen in Table 4.

Table 4. Result from Naive Bayes Performance Model

	True Good	True Excellent	True Bad	Class Precision
Pred. Good	156	47	1	72%
Pred. Excellent	58	12	0	30%
Pred. Bad	11	1	0	0%
Class Recall	69%	39%	0%	

From the table of test results using the Naive Bayes algorithm above, it can be seen that the performance of each class varies greatly. The model performed quite well in identifying samples of the majority class, "Good," with a precision of 72% and a recall of 69%. Despite some errors, the "Excellent" class can also be identified with a precision of 30% and a recall of 39%. However, the "Bad" class performed very badly, with precision and recall all being 0%.

3.5. Decision Tree

After the previous two tests with the K-nearest neighbor algorithm and Naive Bayes, the third classification is done with the decision tree algorithm. The test results of this algorithm are shown in Table 5.

Table 5. Result from Decision Tree Performance Model

	True Good	True Excellent	True Bad	Class Precision
Pred. Good	197	7	0	71%
Pred. Excellent	68	2	0	22%
Pred. Bad	12	0	0	0%
Class Recall	97%	3%	0%	

From the table of test results using the Naive Bayes algorithm above, it can be seen that the performance varies in identifying each class. The model did well in classifying the class "Good," with a high precision of 71% and an excellent recall of 97%. This indicates that most of the positive predictions for the class "Good" were correct, and the model was able to find most of the samples that were actually "Good." However, for the classes "Excellent" and "bad," the very low precision of 22% for "Excellent" and 0% for "bad" and the minimal recall of 3% for "Excellent" and 0% for "bad" indicate the difficulty of the model in distinguishing and recognizing samples of these classes.

3.6. Comparison of KNN, Naive Bayes and Decision Tree Algorithms

Figure 2 shows a comparison of data mining algorithms that use the classification method between KNN, Decision Tree and Naïve Bayes.

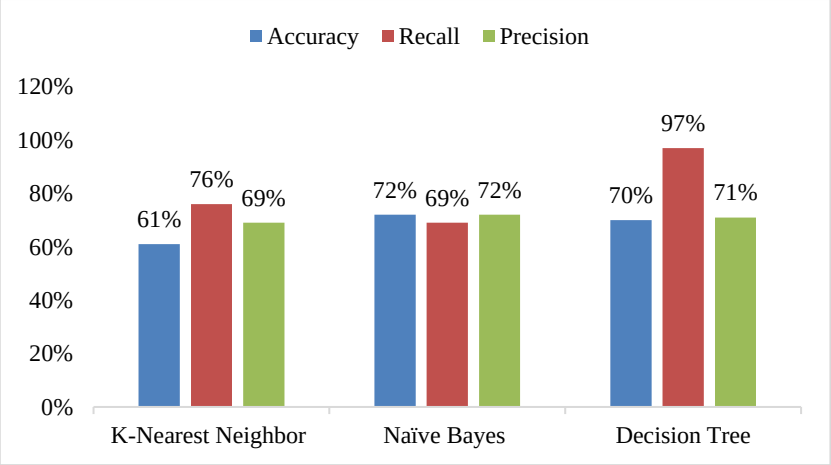

Figure 2. Comparison of KNN, Naive Bayes and Decision Tree Algorithms

The final results of these three classification algorithms show that KNN has a precision of 69%, a recall of 72% and a precision of 61%. A decision tree algorithm with 72% precision has a recall of 69% and a performance precision of 72%. The Naive Bayes algorithm has a precision of 71%, a recall of 97%, and a performance accuracy of 70% for this model. Among these three algorithms, there are quite different levels of comparative accuracy, the KNN algorithm shows a good positive prediction accuracy with 69% accuracy and the ability to identify true positive samples with 72% accuracy, but its accuracy is slightly lower than the other two algorithms. . The decision tree provides a good balance between 72% precision and 69% recall, resulting in a KNN accuracy of 72%. Meanwhile, Naive Bayes stood out with a high recall of 97%, showing its ability to identify almost all true positive samples, albeit with a slightly lower accuracy of 71%. Although Naive Bayes with 70% accuracy shows its superiority in detecting maximum positive samples..

4. CONCLUSION

From the results of the analysis in the research that has been done, based on the classification of movie ratings and TV shows on Netflix, to determine the accuracy of the data, researchers compare the three algorithms. The results of the confusion matrix which includes accuracy, precision, and recall show that the KNN algorithm has an accuracy of 61%, the Naive Bayes algorithm 72%, and the Decision Tree algorithm 70%. Therefore, it can be concluded that the Naive Bayes algorithm has the highest level of accuracy 72%, while the KNN algorithm has the lowest level of accuracy 61%. From these results it can be stated that the Naive Bayes algorithm can classify movie and tv show rating data on Netflix better than compared to the other two algorithms.

REFERENCES

[1] D. W. Azalia and R. H. Magnadi, "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI KEPUTUSAN PEMBELIAN PADA LAYANAN VIDEO ON DEMAND (Studi Pada Pengguna Netflix)," DIPONEGORO JOURNAL OF MANAGEMENT, vol. 9, no. 2, pp. 1–12, 2020, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/dbr

[2] O. Ormanlı, "'Online film platforms and the future of the cinema.' CTC," 2019.

[3] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," Big Data, vol. 1, no. 1, pp. 51–59, Mar. 2013, doi: 10.1089/big.2013.1508.

[4] H. M. Ç.-R. and G. G. Wickham, "R for data science. ' O'Reilly Media, Inc.,'" 2023.

[5] P. N. Harahap and S. Sulindawaty, "Implementasi Data Mining Dalam Memprediksi Transaksi Penjualan Menggunakan Algoritma Apriori (Studi Kasus PT.Arma Anugerah Abadi Cabang Sei Rampah)," MATICS, vol. 11, no. 2, p. 46, Jan. 2020, doi: 10.18860/mat.v11i2.7821.

[6] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[7] X. Song, T. Xie, and S. Fischer, "Accelerating kNN search in high dimensional datasets on FPGA by reducing external memory access," Future Generation Computer Systems, vol. 137, pp. 189–200, Dec. 2022, doi: 10.1016/j.future.2022.07.009.

[8] S. Suyanto, P. E. Yunanto, T. Wahyuningrum, and S. Khomsah, "A multi-voter multi-commission nearest neighbor classifier," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 6292–6302, Sep. 2022, doi: 10.1016/j.jksuci.2022.01.018.

[9] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," Soft comput, vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.

[10] P. Golpour et al., "Comparison of support vector machine, naïve bayes and logistic regression for assessing the necessity for coronary angiography," Int J Environ Res Public Health, vol. 17, no. 18, pp. 1–9, Sep. 2020, doi: 10.3390/ijerph17186449.

[11] E. Ascari, M. Cerchiai, L. Fredianelli, D. Melluso, F. Rampino, and G. Licitra, "Decision trees and labeling of low noise pavements as support for noise action plans," Environmental Pollution, vol. 337, Nov. 2023, doi: 10.1016/j.envpol.2023.122487.

[12] Q. Li, X. Wang, Q. Pei, X. Chen, and K.-Y. Lam, "Consistency preserving database watermarking algorithm for decision trees," Digital Communications and Networks, Jan. 2023, doi: 10.1016/j.dcan.2022.12.015.

[13] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[14] M. Mastur Alfitri and D. Rusda, "Evaluasi Performa Algoritma Naïve Bayes Dalam Mengklasifikasi Penerima Bantuan Pangan Non Tunai," vol. 7, no. 3, pp. 1433–1445, 2023, doi: 10.30865/mib.v7i3.6151.

[15] N. Dalhat Mu'azu and S. Olusanya Olatunji, "K-nearest neighbor based computational intelligence and RSM predictive models for extraction of Cadmium from contaminated soil," Ain Shams Engineering Journal, vol. 14, no. 4, Apr. 2023, doi: 10.1016/j.asej.2022.101944.

[16] M. M. Saritas and A. Yasar, "International Journal of Intelligent Systems and Applications in Engineering Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE, vol. 7, no. 2, pp. 88–91, 2019, doi: 10.1039/b000000x.

[17] A. M. A. K. and A. K. M. Masum. Rahat, "Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset.," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). IEEE, 2019.

[18] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," Comput Chem Eng, vol. 128, pp. 392–404, Sep. 2019, doi: 10.1016/j.compchemeng.2019.06.001.

[19] A. Musadi, C. C. Tertius, J. Steven, H. A. Saputri, and K. M. Suryaningrum, "Comparing Artificial Neural Network and Decision Tree Algorithm to Predict Tides at Tanjung Priok Port," Procedia Comput Sci, vol. 227, pp. 406–414, 2023, doi: 10.1016/j.procs.2023.10.540.

[20] P. Kanani and M. Padole, "Deep learning to detect skin cancer using google colab," Int J Eng Adv Technol, vol. 8, no. 6, pp. 2176–2183, Aug. 2019, doi: 10.35940/ijeat.F8587.088619.

[21] F. R. V. Alves and R. P. Machado Vieira, "The Newton Fractal's Leonardo Sequence Study with the Google Colab," International Electronic Journal of Mathematics Education, vol. 15, no. 2, Dec. 2019, doi: 10.29333/iejme/6440.

[22] S. Ray, K. Alshouiliy, and D. P. Agrawal, "Dimensionality reduction for human activity recognition using google colab," Information (Switzerland), vol. 12, no. 1, pp. 1–23, Jan. 2021, doi: 10.3390/info12010006.