JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1425

IMPLEMENTATION OF DATA MINING ALGORITHM FOR PREDICTING POPULARITY OF PLAYSTORE GAMES IN THE PANDEMIC PERIOD OF COVID-19

Daning Nur Sulistyowati[1*]; Norma Yunita[2]; Siti Fauziah[3]; Risca Lusiana Pratiwi[4]

Information Systems Study Program
Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
www.nusamandiri.ac.id
[1*]daningnur.dgs@nusamandiri.ac.id, [2]norma.nyt@nusamandiri.ac.id, [3]siti.suz@nusamandiri.ac.id,
[4]risca.ral@nusamandiri.ac.id

(*) Corresponding Author

Abstract— The existence of the COVID-19 virus makes everyone fill their time at home by doing various activities, one of them playing games on the phone. For the game to develop continuously, it needs an assessment that comes from the community and especially the game lovers themselves. This assessment is used to find out what category of game you want. Therefore the analysis is needed to determine the interests of game lovers by analyzing the popularity of a game. This research was conducted to predict the level of popularity of games in PlayStore applications to find out how many popular and unpopular games and the accuracy obtained with the C4.5 algorithm and Naive Bayes algorithm. The results obtained using the C4.5 algorithm showed 73 popular games and 12 unpopular games with an accuracy value of 85.83% with a precision of 85.83% and a recall of 100% and Naive Bayes showed 23 popular games and 62 unpopular games with an accuracy value of 80% with a precision of 96.11% and a recall of 81.01%. The evaluation results with the ROC curve show the AUC value using the Naive Bayes model of 0.776 and the C4.5 model of 0.500. Of the two models used, one of them is included in the classification of Good classification, namely the Naive Bayes algorithm model, because it has an AUC value between 0.80-0.90. While the C4.5 algorithm model is included in the Fair classification, has an AUC value between 0.70 - 0.80.

Keywords: C4.5 Algorithm, Naive Bayes Algorithm, Game Popularity, Prediction, Playstore.

Abstrak— *Adanya virus covid-19 membuat semua orang mengisi waktu dirumah dengan melakukan berbagai kegiatan salah satunya bermain permainan dihandphone. Agar game berkembang secara terus menerus dibutuhkannya sebuah penilaian yang berasal dari masyarakat dan khususnya para pecinta game itu sendiri. Penilaian ini digunakan untuk mengetahui kategori game apa yang diinginkan. Maka dari itu dibutuhkan sebuah analisa untuk mengetahui minat dari pecinta game dengan menganalisis popularitas suatu game. Penelitian ini dilakukan untuk memprediksi tingkat kepopularitasan game pada aplikasi playstore dengan tujuan untuk mengetahui berapa banyak game yang populer dan tidak populer da tingkat akurasi yang didapat dengan metode algoritma C4.5 dan naive bayes. Hasil yang didapatkan menggunakan algoritma C4.5 menunjukan 73 game popular dan 12 game tidak popular dengan nilai accuracy sebesar 85,83% dengan precision sebesar 85,83% dan recall sebesar 100% dan naive bayes menunjukan 23 game populer dan 62 game tidak populer dengan nilai accuracy sebesar 80% dengan precision sebesar 96,11% dan recall sebesar 81,01%. Hasil evaluasi dengan ROC curve menunjukan nilai AUC menggunakan model naive bayes sebesar 0,776 dan model C4.5 sebesar 0,500. Dari kedua model yang digunakan satu diantaranya termasuk kedalam klasifikasi Good classification yaitu model algoritma naive bayes, dikarenakan memiliki nilai AUC diantara 0,80 – 0,90. Sedangkan model algoritma C4.5 termasuk kedalam klasifikasi Fair classification, memiliki nilai AUC diantara 0,70 - 0,80.*

Kata Kunci *Algoritma C4.5, Algoritma Naive Bayes, Popularitas Game, Prediksi, Playstore.*

INTRODUCTION

The existence of the covid-19 virus makes the formation of rules implemented by the government such as Large Scale Social Restrictions (PSBB) and Social Distance which affect all activities carried out at home. Work and study must be done at home, even children cannot leave the house to play. To fill the time at home can do various activities one of them playing computer games or on the phone. The game or what we often call the term game is one tangible manifestation of the use of technology. The variety of games available makes it difficult for consumers to determine which games are popular today.

Several previous studies that have been carried out are the Application of K-Means Clustering Algorithm to the Online Battle Arena Multiplayer Game Character [1], Development of the Mahameru Peak Roaming 3D Adventure Game By Determining the Best Climbing Path Using Algorithm A [2], Implementation of Apriori Algorithms for Analysis of Type Selection Characters in Mobile Legend Game [3], Implementation of Fisher Yate Shuffle Algorithm in Educational Games as Learning Media [4], Implementation of Gdlc System Development Model and Linear Congruential Generator Algorithm in Puzzle Games [5], Implementation of Decision Tree C4.5 Algorithm in Selection of Leader Warrior in the Android Power Ranger Legacy Wars Game [6], Implementation of the Fisher-Yates Shuffle and Fuzzy Tsukamoto Algorithm on the Android-Based Adventure Game The Thole Using Game Engine Units [7], Application of Negamax and Alpha Beta Pruning Algorithms in the Othello Game [8], Game Evaluation of Patient Based Shoulder Injury Rehabilitation Patients Movement Ki nect using Kalman Filter [9], Educational Simulation Game for Making SIM Using Neural Network Backpropagation as a Graduation Determination Recommendation [10] and others.

The method of data mining algorithm C4.5 and Naive Bayes is a classification method that is often done in conducting predictive research and has the highest level of accuracy. Based on previous research studies only discuss a game and no research discusses the prediction of game popularity. Therefore this study was conducted to determine the popularity and unpopularity of existing games using the C4.5 algorithm and Naive Bayes algorithm.

This research was conducted to predict the level of popularity of games in PlayStore applications to find out how many popular and unpopular games and the level of accuracy obtained by the C4.5 algorithm and Naive Bayes algorithm.

MATERIALS AND METHODS

Decision Tree

C4.5 algorithm is an algorithm that is commonly used to form a decision tree to predict or classify very strong ones. Decision trees make large data sets into smaller collections using a set of decision rules [11].

Algorithms C4.5 is a development of the ID3 algorithm which is generally used to build a decision tree by doing the following steps, choosing attributes as root, each value is made of branches, each branch has a case, then the process is repeated in all branches so all cases in the branch have the same class [12]. To choose the attribute as root, it is based on the highest gain value of the existing data set. To calculate the gain used a formula like the equation below:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si)$$

[11]..(1)

Information:
S: case set
A: attribute
n: number of attribute attributes A
| Si |: number of cases on the i-th partition
| S |: number of cases in S

You can see the entropy value calculation in the following equation:

$$Entropy(S) = \sum_{i=1}^{n} -pi * \log_2 pi \quad [12] \text{ .... } (2)$$

Information :
S: case set
A: features
n: number of partitions S
Pi: the proportion of Si to S

From the available data, the attributes that are used as determining variables in the formation of decision trees include Categories, Offered By, Game Size, Reviews, Game Flow, Graphics, Control, and Content.

Naïve Bayes

Naïve Bayes is a simple probabilistic classification that calculates a set of probabilities by adding up the frequencies and combined values of a given dataset [13].

Naive Bayes is based on a basic hypothesis if the attribute values are tentatively independent of each other if given output value. One of the advantages of using Naive Bayes is that this method only requires a smaller amount of training data or training data to determine the estimated parameters needed in the classification process. Naive Bayes is known to be far better than most expected algorithms. The equation of the Bayes theorem is [14].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \text{ [13] } \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Di mana :

X	: Data with unknown classes
H	: The data hypothesis is a class Specific
P(H|X)	: H hypothesis probability based on conditions X (posterior probability)
P(H)	: Probability of hypothesis H (prior probability)
P(X|H)	: Probability of X based on the conditions at hypothesis H
P(X)	: Probability of X

Following is the flow of the Naive Bayes method:

1. Start
2. Read the training data
    a. First, calculate P (Ci) for each class
    b. Then calculate P (X | Ci) for each criterion and each class
    c. Then look for P (X | Ci) which is the biggest in conclusion
3. Display prediction results.

Research is a systematic investigation process by studying various materials and sources to find facts, information, and new conclusions. The research method used in this study is the CRISP-DM research method shown in Figure 1.

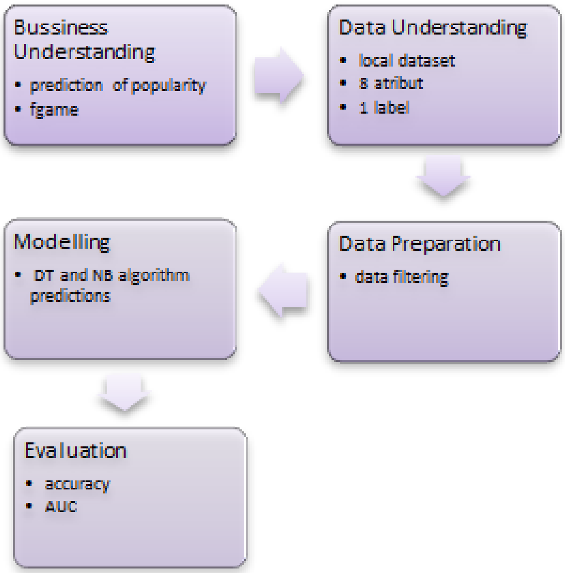


Figure 1. Research Stages

Business Understanding

This stage is commonly referred to as the research understanding stage, by determining the objectives of a research project in the formulation and definition of data mining problems. Based on this data mining classification will produce data classification values based on the algorithm used in predicting the popularity of games in the google play store.

Data Understanding

The understanding data phase includes data collection, conducting data analysis, and evaluating the data. The data source used is the game title data in the play store with 8 attributes. Then the data is analyzed to estimate the amount of data to be retrieved and calculate the amount of data with popular and unpopular information.

Data Preparation

This stage is to prepare data that will be used in the modeling process, data is selected and cleaned as needed. The data used are 85 popular game titles with 8 attributes and 1 label attribute obtained through the Play Store application.

Modeling

Data processing is performed at this stage using data mining techniques. The approaching model used is a Decision Tree C4.5 and Naïve Bayes algorithm approach to producing a predictive value that forms a decision tree.

Evaluation

The evaluation stage is the same as the classification stage where testing is determined for accuracy. The testing phase is carried out to see the results of the application of the proposed algorithm and evaluation using the confusion matrix and the ROC curve. The purpose of this evaluation is to determine the value of the model created in the previous stage.

RESULTS AND DISCUSSION

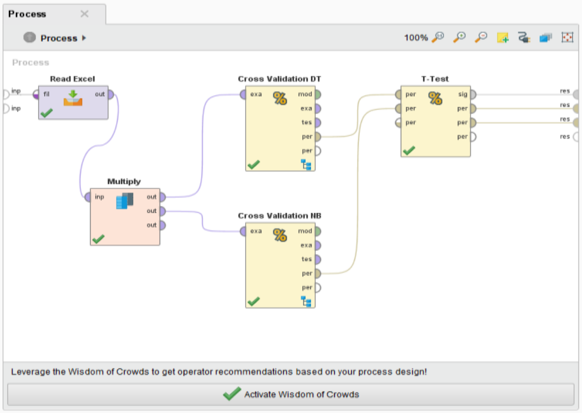
The purpose of this study is to find out how many popular and unpopular games and the accuracy obtained by the C4.5 algorithm and the Naive Bayes method. The dataset is obtained through data collection with conventional features by visiting the website https://play.google.com/store/apps/category/GAME. The data collected were 85 game titles based on the order in which the top 5 appear in each game category. Game data can be seen in table 1.

Tabel 1. Dataset game di Playstore

No	Game Names	Category	Offered by	Games size	Review	Game Flow	Graphic	Control	Content
1	Spiral Roll	Arcade	VOODOO	47MB	72.952	3,4	3,4	3,5	3+
2	Car Master 3D	Arcade	SayGames	73MB	23.703	4,6	4,6	4,5	3+
3	Subway Surfers	Arcade	SYBO Games	129MB	33.162.715	4,3	4,3	4,2	3+
4	Tower Run	Arcade	VOODOO	88MB	346.555	2,7	2,7	2,7	12+
5	Stickman Party: 1 2 3 4 Permainan Pemain Gratis	Arcade	Playmax Game Studio	55MB	364.677	4,4	4,2	4,4	7+
6	Street Racing 3D	Balapan	Ivy	87MB	1.097.536	4,1	4,1	4,0	3+
7	Lomba Sepeda Motor Nyata 3D	Balapan	Italic Games	21MB	624.819	4,0	4,0	3,9	3+
8	Rally Fury - Balap Mobil reli ekstrim	Balapan	Refuel Games Pty Ltd	89MB	309.416	4,3	4,2	4,3	3+
9	Traffic Rider	Balapan	Soner Kara	78MB	7.159.366	4,4	4,4	4,4	3+
10	Hill Climb Racing	Balapan	Fingersoft	76MB	9.735.140	4,3	4,2	4,3	3+

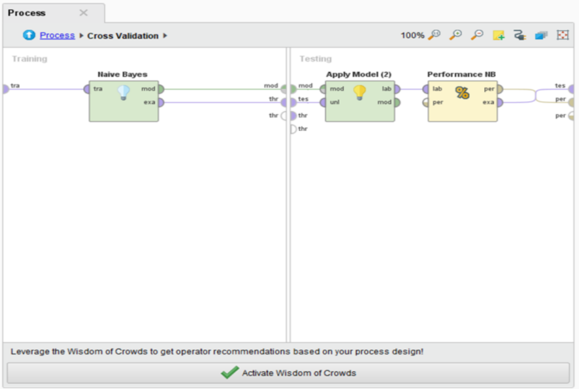
Based on the dataset table above it can be seen that the attributes used are 8 attributes consisting of categories, offered by, game size, reviews, game flow, graphics, controls, and content.

The next step consists of testing and testing the model by calculating the algorithm model used. The algorithm model used in this study is the classification of data mining which consists of C4.5 and Naive Bayes which can be seen in vaults 1 and 2.



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

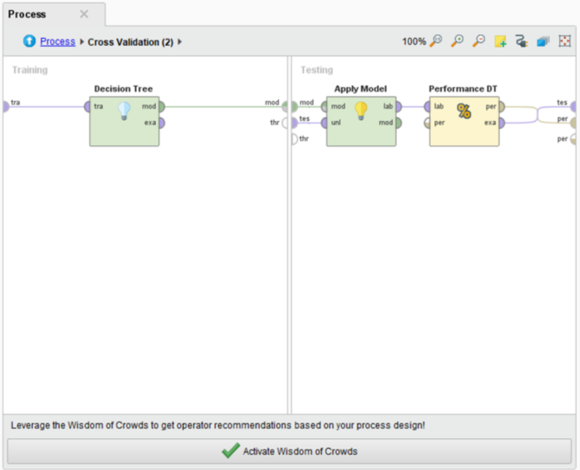
Figure 1. Testing Data Mining with Rapid Miner



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 2. Naive Bayes Algorithm Model Validation

Figure 2 shows the algorithm operator model contained in the validation operator where the case used is the Naive Bayes algorithm model. Based on the test results above it can be seen that the accuracy value for the Naive Bayes algorithm is 80%.



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 3. Decision Tree Algorithm Model Validation

Figure 3 shows the algorithm operator model contained in the validation operator where the case used is the Naive Bayes algorithm model. Based on the test results above, it can be seen that the accuracy value for the Naive Bayes algorithm is 85.83%.

The next step is evaluation using AUC (Area Under Curve). As discussed earlier, the purpose of this study is to predict the level of accuracy of game popularity based on predetermined variables.

The test results using confusion matrix known amount of data following predictions made by the C4.5 algorithm there are 0 data classified as unpopular and 73 data are predicted to be popular, then 0 data is predicted to be unpopular but in C4.5 it is popular and 12 data is predicted to be popular but by C4.5 the algorithm is classified as unpopular.

While the test results using the Naive Bayer method, there are 9 data classified as unpopular and 59 data predicted to be popular. Then 14 data are predicted to be unpopular but Naive Bayes is produced popularly and 3 data are predicted to be popular but by the Naive Bayes algorithm classified as unpopular. So that the formed ROC curve can be seen in Figure 4 and Figure 5.

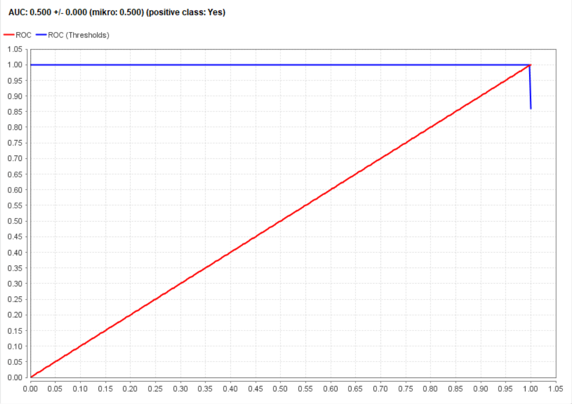


Figure 4 Decision Tree ROC curve

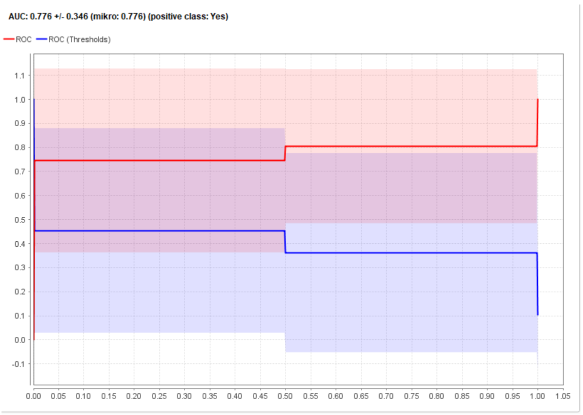


Figure 5. Naive Bayes ROC curve

Based on the results of testing of two data mining classification algorithms that are used show the value of accuracy, prediction, and recall which can be seen in Table 2.

Tabel 2. Nilai Hasil Pengujian

	C4.5	Naive Bayes
Accuracy	85.83%	80%
Precision	85.83%	96.11%
Recall	100%	81.01%

The evaluation results using the ROC curve are obtained through a comparison of the results of the calculation of the AUC value for the classification of data mining used. The evaluation results can be seen in Table 3.

Tabel 3. Nilai AUC

	C4.5	Naive Bayes
AUC	0,500	0,776

AUC evaluation results show the highest value using the Naive Bayes algorithm of 0.776 and the lowest C4.5 value of 0.500. Of the two models used, one of them is included in the classification of Good classification, namely the Naive Bayes algorithm model, because it has an AUC value between 0.80-0.90. While the C4.5 algorithm model is included in the Fair classification, has an AUC value between 0.70 - 0.80.

CONCLUSION

Based on research that has been done using data mining algorithms namely C4.5 and Naive Bayes which are then evaluated with AUC (Area Under Curve) using game data on Playstore to predict game popularity. Confusion Matrix value generated by using the C4.5 algorithm is 85.83% accuracy value with precision is 85.83% and recall is 100% and Naive Bayes is 80% accuracy value with precision is 96.11% and recall is 81, 01%. The evaluation results with the ROC curve show the AUC value using the Naive Bayes model of 0.776 and the C4.5 model of 0.500. While the results for evaluation using the ROC curve for both models produce the highest AUC value using the Naive Bayes algorithm of 0.776 and the lowest value is the algorithm of C4.5 0.500. Of the two models used 1 of them is included in the classification of Good classification, namely the Naive Bayes algorithm model because it has an AUC value between 0.80 to 0.90. While the C4.5 algorithm model is included in the Fair classification, has an AUC value between 0.70 - 0.80.

REFERENCE

[1] M. Mustofa, “Penerapan Algoritma K-Means Clustering pada Karakter Permainan Multiplayer Online Battle Arena,” *J. Inform.*, vol. 6, no. 2, pp. 246–254, 2019.

[2] I. D. Wijaya, R. A. Asmara, and M. Mentari, “Penerapan Algoritma A* Untuk Penentuan Jalur Pendakian Terbaik Pada Game Petualangan 3D,” *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 3, no. 3, pp. 135–142, 2018.

[3] M. S. R. Mustofa, Arina Selawati, Kurani Mega Asteroid, “Implementasi Algoritma Apriori untuk Analisa Pemilihan Tipe

Karakter pada Permainan Mobile Legend," *J. AKRAB JUARA*, vol. 3, no. 1, pp. 130–141, 2017.

[4] F. Fujiati and S. L. Rahayu, "Implementasi Algoritma Fisher Yate Shuffle Pada Game Edukasi Sebagai Media Pembelajaran.," *CogITo Smart J.*, vol. 6, no. 1, p. 1, 2020.

[5] R. A. Krisdiawan, "Implementasi Model Pengembangan Sistem Gdlc Dan Algoritma Linear Congruential Generator Pada Game Puzzle," *Nuansa Inform.*, vol. 12, no. 2, pp. 1–9, 2018.

[6] R. Setiyawan, "Penerapan Algoritma Decision Tree C4 . 5 dalam Pemilihan Leader Warrior pada Game Android Power Ranger Legacy Wars," vol. 6, no. 1, pp. 7–10, 2019.

[7] A. H. Annazili and A. Qoiriah, "Implementasi Algoritma Fisher-Yates Shuffle Dan Fuzzy Tsukamoto Pada Game Petualangan Si Thole Berbasis Android Menggunakan Game Engine Unity," vol. 01, pp. 188–199, 2020.

[8] William, R. Giovanno, and D. Udjulawa, "Penerapan Algoritma Negamax dan Alpha Beta Pruning pada Permainan Othello," *Jatisi*, vol. 2, no. 2, pp. 181–190, 2016.

[9] M. R. Alimansyah, E. C. Djamal, R. Yuniarti, and A. Arif, "Game Evaluasi Gerakan Pasien Rehabilitasi Cedera Bahu Berbasis Kinect menggunakan Kalman Filter," *Semin. Nas. Apl. Teknol. Inf.*, pp. 32–36, 2018.

[10] W. Muhammad, C. Perdana, and A. Qoiriah, "Game Edukatif Simulasi Pembuatan SIM Menggunakan Neural Network Backpropagation Sebagai Rekomendasi Penentu Kelulusan," vol. 01, pp. 217–227, 2020.

[11] J. T. Informasi, I. Sujai, P. Sarjana, T. Informatika, and U. Dian, "SISWA SEKOLAH MENENGAH ATAS DENGAN," vol. 12, no. April, pp. 42–53, 2016.

[12] L. Ariyani, "KAJIAN PENERAPAN MODEL C45 , SUPPORT VECTOR MACHINE (SVM), DAN NEURAL NETWORK DALAM PREDIKSI," vol. 9, no. 1, pp. 72–86, 2016.

[13] M. Siddik and Y. Desnelita, "Penerapan Naïve Bayes untuk Memprediksi Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Akademis," vol. 2, no. 4, pp. 2–6, 2019.

[14] Y. H. Hui *et al.*, "PENERAPAN ALGORITMA NAIVE BAYES UNTUK MEMPREDIKSI JUMLAH PRODUKSI BARANG BERDASARKAN DATA PERSEDIAAN DAN JUMLAH PEMESANAN PADA CV. PAPADAN MAMA PASTRIES. Volume 1.," *J. Mantik Penusa*, vol. 1, no. 2, pp. 16–21, 2017.