International Journal Of Artificial Intelegence Research
Vol 4, No 1, June 2020, pp. 39 - 48
ISSN: 2579-7298

Comparison Analysis of K-Nearest Neighbor and Naïve Bayes in Determining Talent of Adolescence

Yessi Jusman[1,*], Widdya Rahmalina[2], Juni Zarman[2]

[1] Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta, Bantul, Daerah Istimewa Yogyakarta 55183, Indonesia
[2]Department of Informatics Engineering, Faculty of Engineering, Universitas Abdurrab, Pekanbaru, Riau, Indonesia
[1]*yjusman@umy.ac.id*; [2] widdya.rahmalina@univrab.ac.id* [3]
** corresponding author*

ARTICLE INFO

Article history:
Received 25-08-2019
Revised 18-02-2020
Accepted 09-04-2020

Keywords:
Adolescence
Algorithms
Naïve Bayes
K-Nearest Neighbor
Classification
Dataset

ABSTRACT

Adolescence always searches for the identity to shape the personality character. This paper aims to use the artificial intelligent analysis to determine the talent of the adolescence. This study uses a sample of children aged 10-18 years with testing data consisting of 100 respondents. The algorithm used for analysis is the K-Nearest Neigbor and Naive Bayes algorithm. The analysis results are performance of accuracy results of both algorithms of classification. In knowing the accurate algorithm in determining children's interests and talents, it can be seen from the accuracy of the data with the confusion matrix using the RapidMiner software for training data, testing data, and combined training and testing data. This study concludes that the K-Nearest Neighbor algorithm is better than Naive Bayes in terms of classification accuracy.

I. Introduction

Adolescence is a transition period from childhood to adulthood. The nature of childhood is still inherent in him and consideration of maturity has not been fully formed, adolescence in his consideration is still looking for identity to shape his personality character. In this period the child reaches physical maturity and is expected to be accompanied by emotional maturity and social development. Commitment interaction, deep exploration, and reconsideration of commitments with different identity statuses will play a role in their identity search. The extent to which adolescents find a stable identity is closely related to their psychosocial functioning and well-being This period lasts from around 12 years to 20 years[1]

Teenagers are looking for "who I am, what is my role?" In finding their identity, that is knowing their personal needs and goals to be achieved in their lives, then developing teen interests and talents becomes an important issue. In developing their competencies, adolescents still need guidance from parents and the home and school environment. Parental guidance can help in exploring its potential so that it can optimally explore its intelligence. In explaining intelligence, Gardner uses the word talent or talent. Gardner revealed that there are 8 different types of intelligence in each person, namely linguistic intelligence, visual spatial, kinesthetic, musical, intrapersonal, interpersonal, logical-mathematical, and natural.

To determine the interests and talents of children can be known with the help of experts namely child psychologists. But now there is still reluctance from parents to discuss their child with a psychologist, they think they can handle it themselves. In addition, economic factors are a problem with the high cost of consulting psychologists. The number of child psychologists is also not comparable with the rapid population growth. With the development of technology, tests of interest and talent can be helped by technological tools in accordance with the rules of psychology.

Several researches described the talent management problems that can be solved by using data mining techniques,[2]. K Nearest Neighbor (KNN) algorithm for consulting behavioral disorders in

children was performed by [3][4][5]. The classification model performed by classifier can be used in especially talent management and improving talent management with automated competence assessment[6].

Thus, this paper took the case in the field of psychology to determine the interests and talents of children using data mining technology with the k-nearest neighbor method and naïve Bayes who used the previous data to compare the similarity of training data (data previously obtained through psychological tests) and testing data (new data for testing in obtaining results). This discussion will compare the k-nearest neighbor method and naïve Bayes to produce the best method in determining the child's interests and talents[7][8].

II. Research Method

A. Data Collection

The training data used in this study were taken from 350 existing data and testing data were taken from the results of questionnaires given to children aged 10-18 as many as 148 children. Training data and data testing have the same number of attributes obtained from questioner psychology. All datasets will be selected to get the 17 relevant attributes. From 17 attributes, 8 attributes are used as data input to the classification. The target of the classification is interest talent attribute in Table 1.The data attributes used in this study can be seen in Table 1:

Table 1. Data Attributes Training and Testing Data

No.	*Attributes*	*Specification*
1.	NIS/ID	Number Parent Student / ID code if not the students
2.	Name	Surname Child
3.	Place_of_Birth	place of birth date of
4.	Gender	the child Gender
5.	School	Name of the school / None
6.	Address	home child
7.	Phone_number	number
8.	Question_1	Statement of interests and talents
9.	Question_2	Statement of interests and talents
10.	Question_3	Statement of interests and talents
11.	Question_4	Statement of interests and talents
12.	Question_5	Statement of interests and talents
13.	Question_6	Statement of interests and talents
14.	Question_7	Statement of interests and talents
15.	Question_8	Statement of interests and talents of
16.	Interest_talent	the child's talents interests

B. Classification

The classification algorithm used in this paper is naïve bayes that utilizes simple probabilistic in the data mining and K-Nearest Neighbor process which classification is based on analogy, namely comparing testing data with training data that is close to the object in the test data and has similarities with the testing data.

a) Application of K-Nearest Neighbor K-Nearest Neighbor

Algorithm:

i) Determine the parameter K (number of closest neighbors). The parameter K is K = 5.

ii) Calculates the square of the Euclid distance (query instance) of each object against the sample data provided. Euclidean Formula:

$$D(x,y) = \sqrt{\Sigma_{k-1}^{n}(x_k - y_k)^2}$$

Table 2. Training dan Tabel Testing, Tabel Distance

Code	***Language***	***Match***	**Spatial**	***Kinesthetic***	***Music***	***Interpersonal***	***Intrapersonal***	**Naturalis**	***Interests and Talents***
b1	8	11	11	12	12	7	10	12	?
a1	18	16	14	8	13	17	20	15	Intrapersonal

Name: Teguh Anandar

$$a1, b1 = \sqrt{(b1_1 - a1_1)^2 + (b1_2 - a1_2)^2 + (b1_3 - a1_3)^2 + (b1_4 - a1_4)^2 + (b1_5 - a1_5)^2 + (b1_6 - a1_6)^2 + (b1_7 - a1_7) + (b1_8 - a1_8)}$$

$$a1, b1 = \sqrt{(8-18)^2 + (11-16)^2 + (11-14)^2 + (12-8)^2 + (12-13)^2 + (7-17)^2 + (10-20)^2 + (12-15)^2}$$

$$a1, b1 = \sqrt{(-10)^2 + (-5)^2 + (-3)^2 + (4)^2 + (-1)^2 + (-10)^2 + (-10)^2 + (-3)^2}$$

$$a1, b1 = \sqrt{100 + 25 + 9 + 16 + 1 + 100 + 100 + 9}$$

$$a1, b1 = \sqrt{360}$$

$$a1, b1 = 18{,}97$$

iii) Then sort the objects into groups that have the smallest Euclid distance.

iv) Collect Y category (Nearest Neighbor Classification).

v) By using the most majority Nearest Neighbor category, we can predict the calculated query instance value.

b) Application of Naive Bayes

The flow of the Naive Bayes method is as follows:

i) Read training data

ii) Calculate the Amount and probability, but if the numerical data is:

1. Look for probability values by calculating the appropriate amount of data from the same category divided by the data in that category

Interest and Talent Criteria and Probabilities

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)}$$

$$P(C \mid F_1 \ldots F_n) = \frac{P(C)P(F_1 \ldots F_n \mid C)}{P(F_{1 \ldots} F_n)}$$

Calculating class / label

1. Language Probability = $\frac{43}{300}$ = 0.143
2. Mathematical Probability = $\frac{58}{300}$ = 0.193

3. Spatial Probability = $\frac{24}{300}$ = 0.080
4. Kines probability = $\frac{12}{300}$ = 0.040
5. Probability of Music. = $\frac{29}{300}$ = 0.097
6. Probability of Inter = $\frac{57}{300}$ = 0.190
7. Intra-probability = $\frac{46}{300}$ = 0.153
8. Language Probability = $\frac{31}{300}$ = 0.103

Table 3. Training dan Tabel Testing, Tabel Distance

Code	*Language*	*Match*	Spatial	*Kinesthetic*	*Music*	*Interpersonal*	*Intrapersonal*	Naturalis	*Total*
Amount	43	58	24	12	29	57	46	31	300
Probability	0,143	0,193	0,080	0,040	0,097	0,190	0,153	0,103	

iii) Obtain values in the mean table, standard deviation and probability.

In this testing phase, several experimental scenarios will be conducted to determine which classification model is accurate to determine children's interests and talents between the Naïve Bayes and K-Nearest Neighbor methods. The evaluation used is accuracy with the experimental scenario data validation using RapidMiner software, will be discussed further in the results and discussion.

III. Result

After analyzing the data by processing the data that has been obtained, namely training data and testing data. This training data was obtained from psychologist data which was initially a questionnaire that had been processed by a psychologist and interest in his talents was known. While the testing data is data obtained from questionnaires or in the new data set which is filled with children aged 10-18 years.

The data that has been obtained will be analyzed by the K-Nearest Neighbor and Naive Bayes algorithm by conducting comparative analysis in determining children's interests and talents. Data processing is in accordance with the stages of the 2 model algorithm by finding children's interests and talents in the testing data from the questionnaire using previously obtained data from psychologists. The discussion about data processing with the two algorithm models contained in the results has been obtained by processing the data using the data mining process with the K-NN and Naive Bayes algorithm models for testing data. Results of testing data for K-Nearest Neighbor and Naive Bayes. Comparative analysis in determining the interests and talents of children with these two algorithms will determine the value of data accuracy. With the accuracy of the data it can be seen that the algorithmic model of the two methods is accurate in determining children's interests and talents. The following are the results of testing with RapidMiner on training data, data testing, and a combination of training data & data testing.

A. Data Accuracy Testing with K-Nearest Neighbor

a) *Testing training data*

Testing the accuracy of the data performed on training data to obtain value accuracy. Testing the accuracy of training data can be seen in the table as a confusion matrix result of testing the accuracy of the data using RapidMiner software. These training data are data that have been

obtained previously, namely data on interests and talents obtained from expert psychologists and known interests and talents of children.

Table 4. Training dan Tabel Testing, Tabel Distance

Accuracy (%)	***:42,22%***								
Code	***True Intra perso nal***	***True Match***	**True Interp erson al**	***True Lang uage***	***True Spatia l***	***True Music***	***True Kinest hethic***	**True Natural is**	***Class Precision***
Pred. Intrapersonal	7	1	1	2	1	2	0	0	50.00%
Pred. Match	3	12	3	1	4	1	2	3	41.38%
Pred. Intrepersonal	1	0	7	6	0	0	3	2	36.84%
Pred. Language	1	0	0	5	1	1	0	1	55.56%
Pred. Spatial	0	0	0	0	1	0	0	0	100.00%
Pred. Music	1	0	4	1	1	4	1	0	33.33%
Pred. Kinesthethic	0	0	0	0	0	0	0	0	00.00%
Pred. Natural	0	0	1	1	1	1	0	2	33.33%
Class Recall	53.85 %	93.31%	43.75 %	31.25 %	11.11 %	44.44 %	0.00%	25.00%	

The accuracy value obtained by the RapidMiner software with the K-Nearest Neighbor model on training data is worth 42.22%. Based on the ROC Curve, the value of accuracy obtained has a level with a diagnosis of failure or failure.

b) *Testing testing data*

Testing the accuracy of the data performed on testing data to obtain value accuracy. Testing the accuracy of testing data can be seen in the table the confusion matrix results of testing the accuracy of the data using RapidMiner software. This testing data is data that has been obtained from the results of a questionnaire filled with children aged 10-18 years.

Table 5. Confusion Matrix K-NN (Data Testing)

accuracy	***:43,33%***									
	true ?	***true Interp ersona l***	***true Bahas a***	true Mate matika /Logik a	***true Intrap ersona l***	***true Musik***	***true Natur alis***	***true Spasia l***	true Kinest etik	***class precisi on***
pred. ?	0	1	0	1	0	0	0	0	0	0.00%
pred. Interpers onal	0	2	1	1	1	1	0	0	0	33.33 %
pred. Bahasa	0	1	1	0	0	0	0	1	0	33.33 %
pred. Matemati ka/Logik a	0	2	1	8	1	0	1	0	0	61.54 %
pred. Intrapers onal	1	0	0	1	1	0	0	0	0	33.33 %
pred. Musik	0	0	0	0	0	1	0	0	0	100.00 %
pred. Naturalis	0	0	0	0	0	0	0	0	0	0.00%

pred. Spasial	1	1	0	0	0	0	0	0	0	0.00%
pred. Kinestetik	0	0	0	0	0	0	0	0	0	0.00%
class recall	0.00%	28.57 %	33.33 %	72.73 %	33.33 %	50.00 %	0.00%	0.00%	0.00%	

The accuracy value obtained RapidMiner software by the with the model K-Nearest Neighbor on testing data is worth 43.33%. Based on the ROC Curve, the value of accuracy obtained with levels of diagnosis is poor classification or worse classification results. However, for computer-based systems, an accuracy value of <60% is acceptable.

c) *Testing training data and testing data*

Testing the accuracy of the data performed on training& data testing data to obtain value accuracy. Data on training& data testing which carried out testing data accuracy amounted to 400 records consisting of 300 training data and 100 testing data. The training& data testing data are data that have been obtained from data that already exists in psychologists and the results of questionnaires filled with children aged 10-18 years. In the table the confusion matrix results of testing the accuracy of the data using RapidMiner software.

Table 6. Confusion Matrix K-NN (Data Testing)

accuracy	***: 55,00%***									
	true Intrapersonal	***true Mathematica/Logical***	***true Interpersonal***	true Bahasa	***true Spatial***	***true Music***	***true Kinesthetic***	***true Naturalis***	true ?	***class precision***
pred. Intrapersonal	12	2	0	2	4	0	0	0	0	60.00 %
pred. Mathematica/Logical	2	20	2	1	3	0	1	2	1	62.50 %
pred. Interpersonal	1	3	14	5	3	2	0	2	0	46.67 %
pred. Bahasa	2	2	0	8	2	2	1	1	0	44.44 %
pred. Spatial	0	0	0	1	0	0	0	2	0	0.00%
pred. Music	0	0	0	0	1	6	2	0	0	66.67 %
pred. Kinesthetic	0	0	0	0	0	0	1	0	0	100.00 %
pred. Naturalis	0	0	0	0	1	1	0	5	0	71.43 %
pred.?	0	0	0	0	0	0	0	0	0	0.00%
class recall	70.59 %	74.07 %	87.50 %	47.06 %	0.00%	54.55 %	20.00 %	41.67 %	0.00%	

The accuracy value obtained Rapid Miner software with the model K-Nearest Neighbor on training& data testing is worth 55.00%. Based on the ROC Curve, the value of accuracy obtained has a level with a diagnosis of failure or failure.

B. Data Accuracy Testing with Naïve Bayes

a) *Testing training data*

Testing the accuracy of the data performed on training data to obtain value accuracy. Testing the accuracy of training data can be seen in the table as a confusion matrix result of testing the accuracy of the data using RapidMiner software. These training data are data that have been obtained previously, namely data on interests and talents obtained from expert psychologists and known interests and talents of children.

Table 7. Training dan Tabel Testing, Tabel Distance

accuracy	***:46,67%***								
	true Intrapersonal	***true Matematika/Logika***	true Interpersonal	***true Bahasa***	***true Spasial***	***true Musik***	***true Kinestetik***	true Naturalis	***class precision***
pred. Intrapersonal	8	1	0	2	0	0	0	0	72.73 %
pred. Matematika/Logika	4	8	5	4	5	0	2	4	25.00 %
pred. Interpersonal	0	3	9	2	0	0	2	0	56.25 %
pred. Bahasa	0	0	0	6	0	1	0	1	75.00 %
pred. Spasial	0	0	0	0	2	0	0	0	100.00 %
pred. Musik	1	0	2	0	1	6	1	0	54.55 %
pred. Kinestetik	0	0	0	1	0	1	0	0	0.00%
pred. Naturalis	0	1	0	1	1	1	1	3	37.50 %
class recall	61.54 %	61.54 %	56.25 %	37.50 %	22.22 %	66.67 %	0.00%	37.50 %	

The accuracy value obtained RapidMiner software bywith themodel Naive Bayeson training datais worth 46.67%. Based on the ROC Curve, the value of accuracy obtained has a level with a diagnosis of failure or failure.

i) Testing testing data

Testing the accuracy of the data performed on testing data to obtain value accuracy. Testing the accuracy of testing data can be seen in the table the confusion matrix results of testing the accuracy of the data using RapidMiner software. Testing. This data is data that has been obtained from the results of a questionnaire filled with children aged 10-18 years.

Table 8. Training dan Tabel Testing, Tabel Distance

accuracy	***:40,00%***									
Code	***true ?***	***true Bahasa***	true Intrapersonal	***true Matematika/Logika***	***true Naturalis***	***true Spasial***	***true Interpersonal***	true Musik	***true Kinestetik***	***class precision***
pred. ?	0	0	0	0	0	0	0	0	0	0.00%
pred. Bahasa	0	3	1	0	1	1	3	1	0	30.00 %

pred. Intrapersonal	0	0	0	0	0	0	1	0	0	0.00%
pred. Matematika/Logika	2	1	0	5	0	1	0	0	0	55.56 %
pred. Naturalis	0	0	0	0	1	0	0	0	0	100.00 %
pred. Spasial	0	0	0	0	0	0	0	0	0	0.00%
pred. Interpersonal	0	1	1	0	0	3	2	0	1	25.00 %
pred. Musik	0	0	0	0	0	0	0	1	0	100.00 %
pred. Kinestetik	0	0	0	0	0	0	0	0	0	0.00%
class recall	0.00%	60.00 %	0.00%	100.00 %	50.00 %	0.00%	33.33 %	50.00 %	0.00%	

The accuracy value obtained with RapidMinerwith themodel Naive Bayeson testing data is worth 40.00%. Based on the ROC Curve, the value of accuracy obtained with levels of diagnosis is poor classification or worse classification results.

ii) Testing traning data and testing data

Testing the accuracy of the data performed on training& data testing data to obtain value accuracy. Data on training & data testing which carried out testing data accuracy amounted to 400 records consisting of 300 training data and 100 testing data. The training& data testing data are data that have been obtained from data that already exists in psychologists and the results of questionnaires filled with children aged 10-18 years. The following table is a table of confusion matrix results from testing the accuracy of data using RapidMiner software.

Table 9. Training dan Tabel Testing, Tabel Distance

accuracy	***:51,67%***									
Code	***true Intrapersonal***	***true Matematika/Logika***	true Interpersonal	***true Bahasa***	***true Spasial***	***true Musik***	***true Kinestetik***	true Naturalis	***true ?***	***class precision***
pred. Intrapersonal	10	1	5	0	1	0	0	0	1	55.56 %
pred. Matematika/Logika	1	15	3	7	1	2	1	3	0	45.45 %
pred. Interpersonal	3	2	15	2	1	0	1	0	0	62.50 %
pred. Bahasa	2	3	3	9	0	2	0	2	0	42.86 %
pred. Spasial	0	0	0	1	3	0	0	0	0	75.00 %
pred. Musik	0	0	0	0	0	3	0	1	0	75.00 %
pred. Kinestetik	0	0	0	1	0	0	0	0	0	0.00%
pred. Naturalis	2	1	1	1	2	0	0	7	0	50.00 %

pred. ?	0	0	0	1	0	0	0	0	0	0.00%
class recall	55.56 %	68.18 %	55.56 %	40.91 %	37.50 %	42.86 %	0.00%	53.85 %	0.00%	

The accuracy value obtained with RapidMiner with the model Naive Bayes on training & testing is worth 51.67%. Based on the ROC Curve, the value of accuracy obtained has a level with a diagnosis of failure or failure.

After going through the testing and evaluation process, a more accurate model is obtained to measure children's interests and talents. The model is used to evaluate training data or testing data.

Data accuracy test results using RapidMiner software with K-Nearest Neighbor and Naive Bayes models have accuracy values that can be seen in Table 10. Comparison of accuracy values between the two models by testing training data, testing data, and combined training & data testing data as indicated in Table 10:

Table 10. Comparison of Value Data Accuracy

Data	***Accuracy***	
	K-NN	***Naïve Bayes***
Training	42,22%	46,67%
Testing	43,33%	40,00%
Training and Testing	55,00%	51,56%

From Table 10. there can be seen a comparison of two models, namely K-Nearest Neigbor and Naive Bayes in obtaining data accuracy values. Testing the training data of the two models can be compared to the Naive Bayes model which is of higher value than the K-Nearest Neighbor with a value of 46.67%, testing the testing data obtained by the K-Nearest Neighbor model which is higher in value than Naive Bayes with a value of 43.33% , and a combined test of training & data testing data obtained a higher K-Nearest Neighbor model than Naive Bayes with a value of 55.00%. However, the accuracy used for a system is the accuracy of the testing data.

From Table 10, the accuracy of testing data with the K-Nearest Neighbor algorithm worth 43.33% is higher than the accuracy of the testing data with the Naive Bayes algorithm, while the accuracy of training and testing data with the K-Nearest Neighbor algorithm worth 55.00% is higher rather than the accuracy of training and testing data with the Naive Bayes algorithm, so it was concluded that for this study, the K-Nearest Neighbor algorithm is more accurate for classifying children's interests and talents than the Naive Bayes algorithm.

IV. Conclusion

This study determined the interests and talents of children aged 10-18 years with data testing consisting of 100 records using the K-Nearest Neighbor and Naive Bayes algorithms with references from training data that is pre-existing data and obtained accurate algorithms. In knowing the algorithm that is accurate in determining the interests and talents of children, it can be seen from the accuracy of the data with the confusion matrix using RapidMiner software on training data, testing data, and a combination of training data & data testing. This study concluded that the K-Nearest Neighbor algorithm is better than Naive Bayes in terms of classification accuracy.

The suggestions for further research regarding the comparative analysis of K-NN and Naive Bayes in determining children's interests and talents so that this research becomes more developed, namely: This research can be developed by comparing with other algorithms, so that the best algorithm can be determined in determining children's interests and talents. This research can be developed with the aim of determining the interest and talent of early childhood. This research can also be developed using other classification algorithms.

Acknowledgment

Thank you for Universitas Muhammadiyah Yogyakarta for supporting the research.

References

[1] M. Bazmara, S. V Movahed, and S. Ramadhani, "KNN Algorithm for Consulting Behavioral Disorders in Children," *J. Basic Appl. Sci. Res.*, vol. 3, p. 12.
[2] E. Crocetti, "Identity Formation in Adolescence: The Dynamic of Forming and Consolidating Identity Commitments," *Child Dev. Perspect.*, vol. 11, no. 2, pp. 145–150, doi: 10.1111/cdep.12226.
[3] R. Dimitrova, A. Chasiotis, M. Bender, Vijver, and F. J. R., "Collective Identity and Well-Being of Bulgarian Roma Adolescents and Their Mothers," *J. Youth Adolesc.*, vol. 43, no. 3, pp. 375–386, doi: 10.1007/s10964-013-0043-1.
[4] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Towards applying data mining techniques for talent management," in *International Conference on Computer Engineering and Applications, IPCSIT*.
[5] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Intelligent DSS for talent management: a proposed architecture using knowledge discovery approach," in *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*.
[6] J. M. Kirimi and C. A. Moturi, "Application of Data Mining Classification in Employee Performance Prediction," *Int. J. Comput. Appl.*, vol. 146, no. 7, pp. 28–35.
[7] N. Nikitinsky, "Improving Talent Management with Automated Competence Assessment: Research Summary," in *Scientific-Practical Conference" Research and Development-2016*.
[8] A. K. Sharma, "Data Mining Based Predictions for Employees Skill," *Int. J. Adv. Res. Comput. Sci.*, vol. 4, no. 3.