International Journal of Information System & Technology
Akreditasi No. 36/E/KPT/2019 | Vol. 5, No. 1, (2021), pp. 23-30

Case Based Reasoning using K-Nearest Neighbor with Euclidean Distance for Early Diagnosis of Personality Disorder

Anna Hendri Soleliza Jones[1], Cicin Hardiyanti[2]

Department of Informatic, Faculty of Industrial Engineering
Universitas Ahmad Dahlan
Ringroad Selatan, Kampus 4 UAD, Yogyakarta 55191

e-mail: annahendri@tif.uad.ac.id[1], cicin1600018132@webmail.uad.ac.id[2]*

Abstrak

A personality disorder is a condition of a person with an extreme personality that causes the sufferer to have unhealthy and different thoughts patterns and behavior from other people. The personality disorders discussed in this study consisted of 110 diseases with 300 case data and 68 symptoms. Based on Basic Health Research (Riskesdas) 2018 data, it shows that more than 19 million people aged 15 years and over were affected by mental-emotional disorders. Data from the Statistics Indonesia in 2019 that the population of Indonesia is around 265 million people, while according to the Indonesian Clinical Psychologist Association, the number of verified professional psychologists is 1,599 clinical psychologists out of a total membership of 2,078 as of January 2019. However, this figure does not meet the standards of the World Health Organization (WHO), which is that psychologists serve 30 thousand people. This shows that Indonesia still lacks around 28,970 psychologists. The unequal distribution of professional psychologists has made psychologists need a long time to provide a diagnosis because of the number of patients being inversely proportional to the availability of psychologists in Indonesia. Moreover, there is not enough patient knowledge about the symptoms they feel. This study aims to produce a system for diagnosing personality disorders. This study is a case based reasoning to solve problems that have occurred in previous cases using K-Nearest Neighbor to classify data based on the closest distance using the calculation of the Euclidean Distance. Algorithm testing for the system used the Confusion Matrix test. Based on the results of testing data in the 60 case data using K-nearest Neighbor and the calculation of the Euclidean Distance with a score of K=3, it is known that 60 data have 100% similarity to cases with a personality disorder. Meanwhile, testing new cases with 10 case data that were not in the knowledge base was also conducted showing that 9 cases had 100% similarity to the previous case, while another case had 90% similarity to the previous case.

***Kata Kunci**: Case Based Reasoning, K-Nearest Neighbour, Euclidean Distance, Personality Disorder, Diagnosis.*

1. Introduction

A personality disorder is a term used for mental disorders. This condition is described by the existence of an extreme personality that causes sufferers to have different thought patterns and behaviors from normal humans [1]. This behavior usually occurs during adolescence or early adulthood. During this period, a person experiences significant changes biologically, psychologically, and socially. A personality disorder is divided into 10 types, namely paranoid personality disorder, schizoid personality disorder, passive-aggressive personality disorder, antisocial personality disorder, narcissistic personality disorder, histrionic personality disorder, borderline personality disorder, avoidant personality disorder, obsessive-compulsive personality disorder, and dependent personality disorder.

ISSN : 2580-7250

Based on the data from the 2018 Basic Health Research (*Riskesdas*) issued by the Ministry of Health, the prevalence of mental-emotional disorders in people aged > 15 years old in 2013 to 2018 increased from 6& to 9.8% per province [2]. Data from the Statistics Indonesia (*BPS*) in 2019 showed that the Indonesian population is around 265 million people. Meanwhile, according to the Indonesian Clinical Psychologist Association (*Ikatan Psikologi Klinis,* abbreviated as *IPK*), the number of verified professional psychologists is 1,599 clinical psychologists and a total of 2,078 members as of January 2019. However, this number does not meet the standards of the World Health Organization (WHO) where 1 psychologist serves 30 thousand people. This shows that Indonesia still lacks around 28.970 psychologists.

The unequal distribution of professional psychologists indicating that the number of psychologists in Indonesia does not meet the standards of the World Health Organization (WHO). This condition makes psychologists need a long time to provide a diagnosis since the number of patients is more than the number of psychologists. Furthermore, the patient's knowledge of the symptoms they feel is not sufficient.

This study utilized data from a previous study entitled "Case Base Reasoning for the Diagnosis of Personality Disorders by Utilizing Bayesian Probabilistic." The data used were 300 case data and 68 symptom data [3]. Case data were divided into training data and testing data with a ratio of 80% training data and 20% testing data. The method used in this study was Case Base Reasoning (CBR) using K-Nearest Neighbor (KNN). The Case Base Reasoning method was used to solve problems that had occurred in previous cases and then adopted the information and solutions used in previous cases to solve problems in new cases [4]. K-nearest Neighbor was used to classifying data that had the closest distance [5]. This system aims to streamline the performance of psychologists in providing initial diagnosis to patients and to help sufferers to find out solutions for early treatment of personality disorder.

2. Research Methodology

2.1. Case Base Reasoning

Case Base Reasoning (CBR) is a knowledge-based approach to solving problems by utilizing previous experience stored in a case based [6]. The knowledge stored on a case basis reduces the probability of errors occurring and makes it possible to analyze errors in previous cases [7]. A Case based reasoning has diagnostic capabilities that can provide information automatically based on knowledge of previous cases which have been revised according to new case problems. Thus, case base reasoning knowledge can continue to grow to solve problems in the future [8]. In general, there are 4 stages of problem-solving based on case-based reasoning that is used in problem-solving, namely retrieve, reuse, revise, and retain [9]. The problem-solving stage has an important role, which is if you cannot retrieve a case that is similar to the previous case, the CBR stage cannot be continued [10]. The solution is in the form of a cycle as shown in Figure 1, CBR Stages.

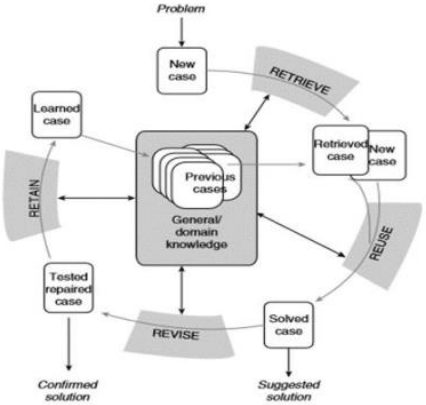

Figure 1. CBR Stages [11]

2.2 K-Nearest Neighbour

K-Nearest Neighbor is a simple classification method to classify new data whose class is not yet known [12]. The new data is classified by finding the closest distance to the object according to the specified number of k [13]. The use of K values is generally an odd number to anticipate the existence of the same distance in the classification process [14]. K-Nearest Neighbor is also known as the lazy learning method because the training sample must be present in memory when the classification process is running [15]. Measurement of the distance to find out the nearest neighbor is calculated using the Euclidean distance, which is a calculation to find the closest distance between two points by determining the similarity of cases to the symptom weight value [16]. The use of the Euclidean distance is useful for determining valid distance [17].

$$\mathbf{d}(\boldsymbol{X_i}, \boldsymbol{X_j}) = \sqrt{\sum_{r}^{n} =_1 (\boldsymbol{a_r}(\boldsymbol{X_i}) - \boldsymbol{a_r}(\boldsymbol{X_j}))^2} \quad (1)$$

Description:
$d\ (\ Xi, Xj)$ = Euclidean Distance
Xi = i record
Xj = j record

2.3 Confusion Matrix

Classification model testing is done to find out how well the system performs data classification by predicting true and false objects. A confusion matrix is used to measure the performance of the classification method [18].

Table 1. Confusion Matrix

Classification	Prediction Class	
	Positive Classification	Negative Classification
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Table 4 presents that the number of positive data classified as correct by the system is called TP while negative data classified correctly by the system is called TN. FP is the number of positive data classified incorrectly by the system, while FN is the number of negative data classified incorrectly by the system [19].

a) *Accuracy* is a test to find out how accurate the system is to classify data correctly. [20]. The accuracy value is the total of all data assessed and identified [21].
Accuracy = TP+TN/TP+TN+FP+FN)*100% (2)
b) *Precision*, the number of positive data classified correctly is then divided by the total data classified as positive [22].
Precission = TP/(TP+FP)*100% (3)
c) *The recall* is used for the proportion of documents that the system can recover [23].
Recall = TP/(FN+TP)*100% (4)

3. Result and Discussion

The system built required case data, disease data, and symptoms for early diagnosis of personality disorder then processed into data that were ready to be implemented into the system. Symptom data consisted of 68 symptoms and disease data consisted of 10 disease data.

Table 2. Disease Data

No	Id	The Name of the Disease
1	PPD	Paranoid Personality Disorder
2	SPD	Schizoid Personality Disorder
3	APD	Antisocial Personality Disorder
4	HPD	Histrionic Personality Disorder

No	Id	The Name of the Disease
5	OPD	Obsessive-Compulsive Personality Disorder
6	AVPD	Avoidant Personality Disorder
7	DPD	Dependent Personality Disorder
8	NPD	Narcissistic Personality Disorder
9	PAPD	Passive-Aggressive Personality Disorder
10	BPD	Borderline Personality Disorder

3.1. Decision Flow

In decision making, a flow was made to describe the flow of decision making using the case base reasoning method. The case-based reasoning stage consisted of 4 steps, namely retrieve, reuse, revise, and retain. Data classification used the k-nearest neighbor method and distance measurement used Euclidean distance. The calculation of k-nearest neighbors used 300 case data including 240 training data and 60 testing data.

Table 3. Training Data

Id	Name	Disease	G1	G2	G3	G4	G5	G68
1	Case 1	1	2	2	3	0	0	0
2	Case 2	1	2	0	3	2	2	0
3	Case 3	1	0	0	3	0	0	0
4	Case 4	1	2	0	3	0	0	0
5	Case 5	1	2	0	3	0	0	0
…	…	…	…	…	…	…	…	…
296	Case 296	10	0	0	0	0	0	0
297	Case 297	10	0	0	0	0	0	0
298	Case 298	10	0	0	0	0	0	0
299	Case 299	10	0	0	0	0	0	0
300	Case 300	10	0	0	0	0	0	2

Table 4. Testing Data

Id	Name	Disease	G1	G2	G3	G4	G5	G68
1	Patient 1	1	2	2	3	0	0	0
2	Patient 2	1	2	0	3	2	2	0
3	Patient 3	1	0	0	3	0	0	0
…	…	…	…	…	…	…	…	…
238	Patient 238	10	0	0	0	0	0	0
239	Patient 239	10	0	0	0	0	0	0
240	Patient 240	10	0	0	0	0	0	2

After the data were divided into training data and testing data, calculations were performed using the k-nearest neighbor method using the Euclidean distance calculation.

3.2. K-Nearest Neighbor Calculation

K-nearest neighbor calculation was done to find the nearest neighbor between cases by using the Euclidean distance calculation.

Table 5. Calculation of Euclidean Distance

id	Disease	Euclidean Distance
1	1	2.83
2	1	2.83
3	1	2.83
…	…	…
238	10	7.81
239	10	7.81
240	10	7.81

After calculating the distance using Euclidean distance, the next step was to sort the closest distance based on the minimum distance K=3. Distances with the same value were sorted into the same class. The distance that satisfied K=3 was disease 1, namely paranoid personality disorder. The next step was to determine the categories included in K=3. Category values were taken based on the closest distance, namely distance 1, 2, and 3. Distances that were more than 3 were not included in the K=3 category.

Table 6. Category Determination

No	Disease	Euclidean Distance	Order of Distance	KKN = 3	Yes Category for KNN
1	1	2.83	2	Yes	Paranoid
2	1	2.83	2	Yes	Paranoid
3	1	2.83	2	Yes	Paranoid
…	…	…	…	…	…
238	10	7.81	8	No	-
239	10	7.81	8	No	-
240	10	7.81	8	No	-

Based on the Yes category table for K-NN in column 6, there were 35 cases of paranoid personality disorder. After the categories had been determined, the next step was to determine the classification. The classification was taken based on the majority of the disease. Thus, the results were obtained, namely paranoid personality disorder with 35 cases.

3.3. Testing of New Cases

Testing of new cases was conducted using data not found in training data or testing data. The data used in new cases was 10 cases.

a) Calculation of new Euclidean distance

Table 7. New Case Euclidean Distance Calculation

id	Disease	Euclidean Distance
1	1	6.24
2	1	6.24
3	1	6.86
…	…	…
298	10	7.48
299	10	6.63
300	10	6.93

b) Determined the closest distance

The distance determination was based on the minimum value of K, namely K=3. The distance taken was the distance that met the K value, namely schizoid personality disorder, paranoid personality disorder, and avoidant personality disorder.

c) Define the category

Table 8. Determination of New Case Categories

No	Disease	Euclidean Distance	Order of Distance	KKN = 3	Yes Category for KNN
1	1	6.24	2	Yes	Paranoid
2	1	6.24	2	Yes	Paranoid
3	1	6.86	6	No	-
…	…	…	…	…	…
298	10	7.48	10	No	-
299	10	6.63	4	No	-
300	10	6.93	7	No	-

Based on the Yes category table for K-NN in column 6, there were 12 cases of paranoid personality disorder. After the categories have been determined, the next step was to determine the classification based on the majority of diseases. The results of the classification of the test data were paranoid personality disorder with a total of 12 cases.

3.4. Algorithm Testing of the System

Accuracy testing is an accuracy test performed to measure the performance of the classification method [24]. This test used the Confusion Matrix by calculating accuracy, precision, and recall.

a) Testing of Testing Data

The results of accuracy testing with confusion matrix for 20% testing data, 6 case data can be seen in table 7 of the confusion matrix data testing.

Table 9. Testing Data Confusion Matrix Testing

Classification	Prediction Class	
	Positive Classification	Negative Classification
Positive	60 TP	0 FP
Negative	0 FN	0 TN

Testing the system with the Confusion Matrix using the Euclidean Distance method showed that 60 cases had similarities with a personality disorder. The results of the calculation were 100% accuracy, 100% precision, and 100% recall.

b) Testing of New Case Test Data

The results of accuracy testing with confusion matrix for 10 case data are presented in table 8, testing confusion matrix data test.

Table 10. New Case Confusion Matrix Testing

Classification	Prediction Class	
	Positive Classification	Negative Classification
Positive	9 TP	0 FP
Negative	1 FN	0 TN

4. Conclusion

This study has implemented case based reasoning with the K-Nearest Neighbor and Euclidean Distance methods which aim to produce a system that can diagnose personality disorders. K-Nearest Neighbor is used to classify data based on the closest distance using Euclidean Distance calculations. Based on the results of data testing on 60 case data using K-Nearest Neighbor and the calculation of Euclidean Distance with a score of K = 3, it is known that 60 data has 100% similarity to cases of personality disorder. Meanwhile, testing of new cases with 10 case data which were not in the knowledge base was also carried out which showed that 9 cases had 100% similarity to the previous case, while other cases had 90% similarity with the previous cases. System testing with the Confusion Matrix using the Euclidean Distance method shows 9 cases have similarities with personality disorders and are positive for actual and predicted values. Meanwhile, 1 case is positive for the actual value and negative for the prediction. Calculation results with 90% accuracy, 100% precision, and 90% recall.

References

[1] L. Ekselius, “Personality disorder: a disease in disguise,” *Ups. J. Med. Sci.*, vol. 123, no. 4, pp. 194–204, 2018.

[2] K. Riskesdas, “Hasil Utama Riset Kesehata Dasar (RISKESDAS),” *Has. Utama Riskesdas 2018*, vol. 44, no. 8, pp. 1–200, 2018.

[3] A. Hendri Soleliza Jones, "Case Based Reasoning Untuk Diagnosis Gangguan Kepribadian Dengan Memanfaatkan Probabilistic Bayesian," vol. 1, no. 1, 2016.

[4] Y. S. Bagi and S. Suprapto, "Hepatitis Diagnosis Using Case-Based Reasoning with Gradient Descent as Feature Weighting Method," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 4, no. 1, p. 25, 2018.

[5] V. B. Surya Prasath, H. A. Abu Alfeilat, O. Lasassmeh, A. B. A. Hassanat, and A. S. Tarawneh, "Distance 1–39, 2017.

[6] H. Santoso and A. Musdholifah, "Case Base Reasoning (CBR) and Density Based Spatial Clustering Application with Noise (DBSCAN)-based Indexing in Medical Expert Systems," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 5, no. 2, pp. 169–178, 2019.

[7] S. Tahmasebian, M. Langarizadeh, M. Ghazisaeidi, and M. Mahdavi-Mazdeh, "Designing and implementation of fuzzy case-based reasoning system on android platform using electronic discharge summary of patients with chronic kidney diseases," *Acta Inform. Medica*, vol. 24, no. 4, pp. 266–270, 2016.

[8] E. P. Purwandari, A. P. Yani, R. Sugraha, K. Anggriani, and E. Widi Winarni, "Online Expert Systems For Bamboo Identification Using Case Based Reasoning," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 5, p. 2766, 2017.

[9] M. Benamina, B. Atmani, and S. Benbelkacem, "Diabetes Diagnosis by Case-Based Reasoning and Fuzzy Logic," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 3, p. 72, 2018.

[10] Z. Zhai, J. F. M. Ortega, N. L. Martínez, and H. Xu, "An efficient case retrieval algorithm for agricultural case-based reasoning systems, with consideration of case base maintenance," *Agric.*, vol. 10, no. 9, pp. 1–21, 2020.

[11] P. Zhang, A. Essaid, C. Zanni-Merk, and D. Cavallucci, "Case-based Reasoning for Knowledge Capitalization in Inventive Design Using Latent Semantic Analysis," *Procedia Comput. Sci.*, vol. 112, pp. 323–332, 2017.

[12] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technol. Heal. Care*, vol. 26, no. S1, pp. S509–S519, 2018.

[13] Okfalisa, R. Fitriani, and Y. Vitriani, "The comparison of linear regression method and k-nearest neighbors in scholarship recipient," *Proc. - 2018 IEEE/ACIS 19th Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distributed Comput. SNPD 2018*, pp. 194–199, 2018.

[14] E. N. dkk Shofia, "Sistem Pakar Diagnosis Penyakit Demam : DBD , Malaria dan Tifoid Menggunakan Metode K-Nearest Neighbor – Certainty Factor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 5, pp. 426–435, 2017.

[15] D. Dananjaya, I. Werdiningsih, and R. Semiati, "Decision Support System for Classification of Early Childhood Diseases Using Principal Component Analysis and K-Nearest Neighbors Classifier," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 5, no. 1, p. 13, 2019.

[16] O. Olaide and A. F. D. Kana, "OWL Formalization of Cases : An Improved Case-Based Reasoning in Diagnosing International Journal of Information Security , Privacy and Digital Forensics OWL Formalization of Cases : An Improved Case-Based Reasoning in Diagnosing and Treatment of Breast Cancer," no. September 2020, 2019.

[17] M. Hoffmann and F. Noé, "Generating valid euclidean distance matrices," *arXiv*, 2019.

[18] P. Singh, S. Singh, and G. pandi jain, "Effective heart disease prediction system using data mining techniques." International Journal Of Nanomedicine, Gujarat, India, p. 4, 2021.

[19] D. Chicco, N. Tötsch, and G. Jurman, "The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and

markedness in two-class confusion matrix evaluation," *BioData Min.*, vol. 14, pp. 1–22, 2021.

[20] T. Rosandy, "PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4.5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus : KSPPS / BMT AL-FADHILA," *J. Teknol. Inf. Magister Darmajaya*, vol. 2, no. 01, pp. 52–62, 2016.

[21] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 879, no. 1, 2020.

[22] Subari and Ferdinandus, "Sistem Information Retrieval Layanan Kesehatan Untuk Berobat Dengan Metode Vector Space Model (Vsm) Berbasis Webgis," *Snatika2*, no. November, pp. 202–212, 2015.

[23] M. Martin and L. Nilawati, "Recall dan Precision Pada Sistem Temu Kembali Informasi Online Public Access Catalogue (OPAC) di Perpustakaan," *Paradig. - J. Komput. dan Inform.*, vol. 21, no. 1, pp. 77–84, 2019.

[24] D. A. Fauziah *et al.*, "Klasifikasi Berita Politik Menggunakan Algoritma K-nearst Neighbor (Classification of Political News Content using K-Nearest Neighbor) Abstrak," *J. Sist. Inf. Univ. Jember*, vol. 6, no. 2, p. 8, 2018.