Application of the K-Nearest Neighbors (K-NN) Algorithm for Classification of Heart Failure

Ryan Yunus [*1], **Uli Ulfa** [2]
Faculty of Informatics / Pati College of Engineering, Jln Raya Pati Trangkil KM 4,5 Pati, Central Java
*E-mail : riyanyunus@sttp.ac.id[*1], uliulfa@stttp.ac.id[2]*

Melina Dwi Safitri [3]
Faculty of Informatics / Pati College of Engineering, Jln Raya Pati Trangkil KM 4,5 Pati, Central Java
E-mail : melina18dwi @ gmail.com[3]

Abstract - Heart failure is a type of disease that has the largest number of patients in the world. Based on information from the data center, there were 229,696 people with heart failure in 2013. Lack of public knowledge about what indications of a person having heart failure make the main cause not handled properly by heart failure patients. In this study, data classification was carried out using KNN algorithm because it has a simple calculation and has a fast time. This study only uses 12 attributes, while the previous study compared 6 algorithms with 13 attributes from 299 data. The highest algorithm with 94.31% accuracy by Random Forest while KNN had an accuracy rate of 86.95% with the same data. In this study, the accuracy of the sample data was compared between 20 data and 299 total data. Both of them have different accuracy. 20 sample data has an accuracy rate of 89.29% while 299 data has an accuracy rate of 96.66%.

Keywords - RapidMiner, classification, k-nearest neighbors (KNN), heart failureclinic records

1. INTRODUCTION

Heart failure is the number one deadly disease in the world, if not handled properly it will cause death [1]. Heart failure causes hospital admissions to increase by approximately 6.5 million per year [2].

According to doctor's diagnosis at the Indonesian Center for Data and Information, heart failure affected approximately 530,068 people in 2013 [3]. Heart failure has a high potential in the future. A slowing heart rate indicates worsening heart failure [4]. Heart failure sufferers will easily get tired and short of breath when doing activities [5]. A person with heart failure has several differences from other heart sufferers, namely shortness of rest and activity, also easy fatigue [5].

Lack of public knowledge about heart disease is the highest factor causing the increase in heart failure sufferers [6]. This is because people do not know the symptoms of heart failure, so they do not take further treatment in overcoming the disease. The severity of the disease depends on the patient's quality of life. Many analyzes on the quality of the sufferers were carried out so as to reduce heart failure based on physical abilities and disease duration [7].

Heart failure is a condition in which the heart is unable to flow / pump blood to the tissues in the body, while the flow to the heart is still running with a fairly high intensity [8], [9]. Therefore, the researchers propose to classify heart failure.

The purpose of this research is that the public can find out what the indications are for heart failure. Researchers will use the k-nn (k-nearest neighbors) algorithm because this algorithm has the simplest problem solving [10], [11]. In the previous research, the researcher made a prediction of failure disease by comparing 6 algorithms, namely Random Forest, Decission Tree, KNN, SVM, ANN and Naïve Bayes. The highest accuracy results using RF with a combination of sample testing techniques, resulting in an accuracy of 94.31% [9]. Meanwhile, in this study, researchers will use KNN for the classification process.

2. RESEARCH METHOD

2.1. The used data

Data were taken from UCI Machine Learning. The data used has a total of 299 records with 12 attributes. Consisting of age, anemia, diabetes, ejection fraction, high blood pressure, plateles, serum creatine, serum sodium, sex, smoking, time.

Table 1. Data *Heart Failure Clinical Process*

no	Age	anaemia	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
1	75	0	0	20	1	265000	1.9	130	1	0	4	1
2	55	0	0	38	0	263358	1.1	136	1	0	6	1
3	65	0	0	20	0	162000	1.3	129	1	1	7	1
4	50	1	0	20	0	210000	1.9	137	1	0	7	1
5	65	1	1	20	0	327000	2.7	116	0	0	8	1
….	90	1	0	40	1	204000	2.1	132	1	1	8	1
…..	75	1	0	15	0	127000	1.2	137	1	0	10	1
299	75	1	0	15	0	127000	1.2	137	1	0	10	1

2.2. Classification

Classification is a process to describe and distinguish certain classes in order to be used in predicting data based on labels from classes whose labels are not yet known [12]. There are several classification algorithms used, namely K-nn, Decision Tree, Naïve Bayes, etc. Some of these algorithms have different calculation and prediction methods according to their function. The classification process is used to determine the values that often appear based on K objects [13].

2.3. K-NN (K-Nearest Neighbors)

K-Nearest Neighbors is a simple data mining algorithm to solve problems in classification [10]. K-nn includes instance-based learning [12]. instance-based learning is also called lazy learn because it is used for learning in machine learning [14].

Yusra stated that the results of the classification process using Knn can be seen from the classes that often appear based on the number of k distance to the nearest / closest neighbor [15]. The stages of the k-nn process are [16]:

1. Enter the k value to be used.
2. Enter the data to be classified.
3. Select a random number of k class centers from the data
4. Calculate the distance that approaches the data member
5. The data with the closest distance will be selected as a prediction.

K-nn processes by finding the distance between two points, testing and training. then the distance equation process will be carried out. The distance equation is called the Euclidean distance. The equation of the Euclidean distance is in equation 1 [13].

$$d_{(x,y)} = \sqrt{\sum_{r-1}^{n} (x_x - x_y)^2} \qquad (1)$$

d (x, y) : *Euclidean distance*
x_i : training data
x_j : test

2.4. The Accuracy of calculation data

After the data through the training and testing process, the data accuracy will be calculated. The accuracy process is a process to determine the level of accuracy and performance of the dataset [17]. Accuracy calculations need to be done for determining the final result. The accuracy of the process results will be calculated using equation 2 [18].

$$Accuracy = \frac{Amount\ of\ correct\ data}{Total\ number\ of\ data} x100 \qquad (2)$$

2.5. The System Design

The system is created using the Rapidminer application. Rapidminer is a data mining simulation application designed based on blocks called operators which have several actions and processes, making the process easier [19].

The system design is made by first looking for data and then entering it in rapidminer. In Rapid Miner, data is processed in classification using the KNN algorithm. The system planning flow is shown in Figure 1.

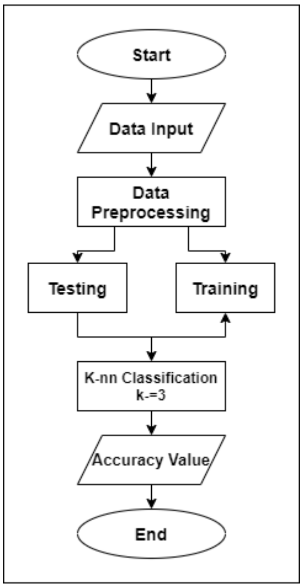

Figure 1. *flowchart* system planning

3. RESULTS AND DISCUSSION

Heart failure clinical record data used in this study amounted to 299. The data were obtained from Uci Machine Learning with 13 attributes.

3.1 Implementation of the algorithm K-Nearest Neighbors

To calculate manually, researchers took 20 data samples from the 299 data used. After the sample data is determined, it is followed by the calculation of the Euclidien distance for each data. The 20 data is divided into testing and training, for the test data used is the 1st data shown in table 2.

Table 2. Data test

No	Age	Anemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
1	75	0	582	0	20	1	265000	1.9	130	1	0	4	1

Data test is data used to test the accuracy of all data The number of k used is 3. K = 3 is used because k = 3 has the optimal result. If the closest distance to the training data is little, then the probability does not have another possibility to enter another class. The higher the k will decrease [20]. Calculating the *Euclidien distance:*

$$d_{(x,y)} = \sqrt{\sum_{r-1}^{n}(x_x - x_y)^2} \quad (3)$$

$$(d1,2) = \sqrt{(55-75)^2 + (0-0)^2 + \cdots\,..+(6-4)^2}$$

(d1,2) = 7461,95

The process is carried out sequentially from the first data to the 21st data. The results of Euclidien Distance and rank are shown in table 3.

Table 3. Results of Euclidien Distance and rank

No	age	anaemia	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT	Prediksi
1	75	0	0	20	1	265000	1.9	130	1	0	4	1	1
2	55	0	0	38	0	263358	1.1	136	1	0	6	1	1
3	65	0	0	20	0	162000	1.3	129	1	1	7	1	1
4	50	1	0	20	0	210000	1.9	137	1	0	7	1	1
5	65	1	1	20	0	327000	2.7	116	0	0	8	1	1
6	90	1	0	40	1	204000	2.1	132	1	1	8	1	1
7	75	1	0	15	0	127000	1.2	137	1	0	10	1	1
8	60	1	1	60	0	454000	1.1	131	1	1	10	1	1
….	…..	…	…	…	…	….	….	….	….	…	….	…	…
….	…..	…	…	…	…	….	….	….	….	…	….	…	…
20	48	1	1	55	0	87000	1.9	121	0	0	15	1	1

Table 3 shows that using the 1st data as testing will produce the euclidien distance value according to the calculations carried out. after the data is calculated the data will be ranked by looking for the nearest 3 value. The 3 closest values can be seen in table 4.

Table 4. Results of closest k

no	age	anemia	….	Death Event	Ecluidien	rank
16	**82**	**1**	**….**	1	**218000,097**	**1**
8	60	1	….	1	189000,1935	2
20	48	1	….	1	178000.0061	3

To predict whether the results of the data from testing 1 include patients who died or not, see table 4. The three closest values are death event 1 (yes). The calculation process is carried out until the 20th sample data.

Table 5. Prediction results

no	age	Anemia	Diabetes	...	Death Even	prediction
1	75	0	0	..	1	1
2	55	0	0	...	1	1
3	65	0	0	...	1	1
4	50	1	0	....	1	1
5	65	1	1	....	1	1
6	90	1	0	....	1	1
7	75	1	0	....	1	1
8	60	1	1	...	1	1
9	65	0	0	....	1	1
10	80	1	0	....	1	1
11	75	1	0	....	1	1
12	62	0	0	....	1	1
13	45	1	0	....	1	1
14	50	1	0	....	1	1
15	49	1	0	....	0	1
16	82	1	0	....	1	1
17	87	1	0	....	1	1
18	45	0	0	....	1	1
19	70	1	0	....	1	1
20	48	1	1	....	1	1

The results of the Euclidean distance from the calculation and reduction are shown in table 5. In calculating the accuracy, the researcher calculates the true and false data.

$$\frac{20-1}{20} x\ 100 \ = \ 95\% \qquad (5)$$

3.4. Implementation of KNN in Rapid Miner

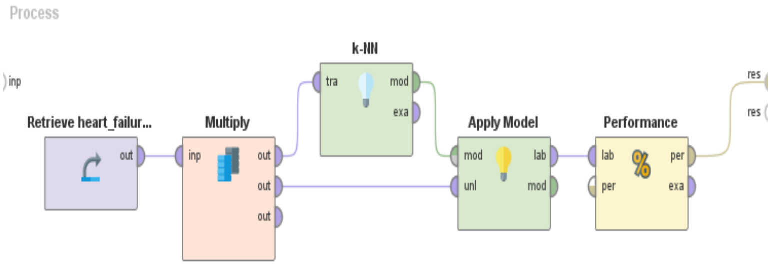

Figure 2. Design

For implementation in Rapid Miner, the researcher conducted two experiments for sample data and population data. The process design of the rapidminer is shown in Figure 2. The next step is the accuracy process of sampling data and population data used. The accuracy results of the sampling data can be seen in Figure 4, which has an accuracy value of 95% and has the same error location in the 15th data sample.

Table 6. results of data sampling accuracy in rapidminer

Accuracy : 89.29 %			
	True1	True0	Class precision
pred1	25	3	89.29%
Pred0	0	0	0.0%
class recall	100%	0%	

Table 6 shows the accurasy of 20 data samples. it has an accuracy of 89.29%. The accuracy of 299 data is shown in table 7. It has an accuracy of 96.66%. From the two tables, it shows that 299 data has a higher accuracy than 20 data.

Table 7**.** the results of the accuracy of all data.

Accuracy : 96.66 %			
	True1	True0	Class precision
Pred 1	96	10	90.57%
Pred 0	0	193	100%
class recall	100%	95%	

Comparison of the results of data accuracy from tables 6 and 7 can be seen in Figure 1 below. Table 6 consists of 20 sampling data, while table 5 consists of all data, namely 299 data. Figure 3 shows that the accuracy data of 299 data from table 7 has a higher accuracy than data of 20.

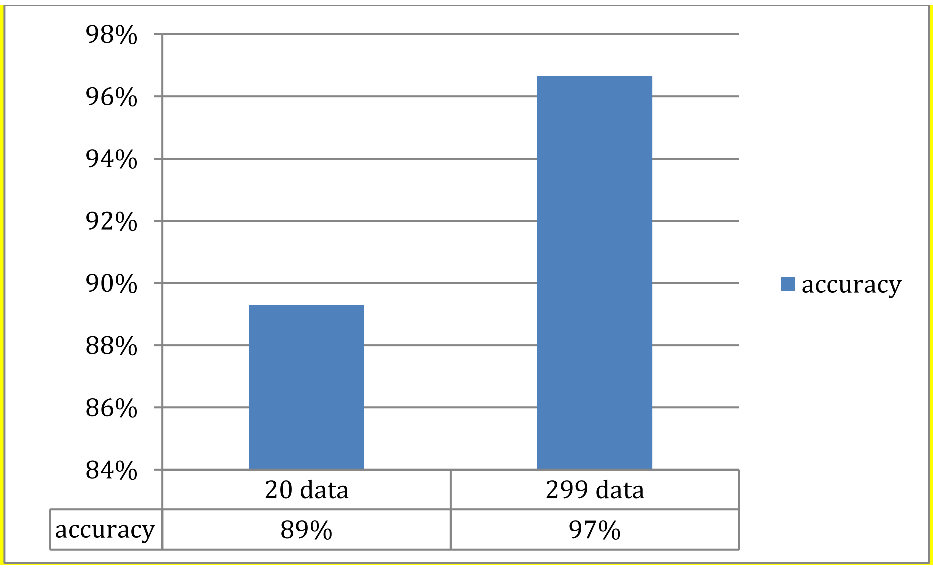

Figure 3. Comparison of accuracy between 20 dan 299 data

4. CONCLUSION

From the heart failure data, it can be concluded that the Knn has high accuracy even with a simple and fast calculation. The accuracy obtained from the data is 96.66%, compared with previous studies, with 13 attributes and now only using 12 attributes. The level of accuracy in this study has increased from 86.95% to 96.66%. These data shows a better degree of accuracy than 13 attribute. By not doing resample tests, the result would be higher and only uses 12 variables. For further research, validation can be done more than once. It can be improved in terms of process, not only for classification.

REFERENCES

[1] Wahyudi E, Hartati S. P. D. dan I. Kementerian Kesehatan RI, "Case-Based Reasoning untuk Diagnosis Penyakit Jantung," *IJCCS (Indonesian J. Comput. Cybern. Syst).*2017; 11(1): 1.

[2] Puspita D, Fadil M. Penggunaan Ventilasi Mekanik pada Gagal Jantung Akut. *J. Kkes. Andalas*. 2020; 9(1S) : 194–203.

[3] P. D. dan I. Kementerian Kesehatan RI, "Situasi Kesehatan Jantung ; Mari Menuju Masa Muda Sehat, Hari Tua Nikmat Tanpa PTM dengan Perilaku Cerdik," *Pus. Data dan Inf.*, p. 8, 2014..

[4] Tanai E, Frantz S. Pathophysiology of heart failure. *Compr. Physiol.* 2016; 6(1): 187–214.

[5] Palilati N. H, Wantania FEN, and Rotty LWA. Hubungan Performa Fisik dengan Prognosis Pasien Gagal Jantung. ECL (JURNAL E-CLINIC) .2021; 9 (28): 118–123.

[6] Putra PD and Rini DP Prediksi Penyakit Jantung dengan Algoritma Klasifikasi. *Pros. Annu. Res. Semin. 2019*; 5(1): 978–979.

[7] Haryati H, Saida S, Rangki L. Kualitas Hidup Penderita Gagal Jantung Kongestif Berdasarkan Derajat Kemampuan Fisik Dan Durasi Penyakit. *Faletehan Heal. J.* 2020; 7(2): 70–76.

[8] Sudoyo A, Setiyohadi B, Alwi I, Simadibrata M K, Setiati. *Buku Ajar llmu Penyakit Dalam.* 5th ed. 2009: .

[9] Rahayu S, Purnama J J, Pohan A B, Nugraha F S, Nurdiani S, Hadianti S. Prediction of survival of heart failure patients using random forest. *J. Pilar Nusa Mandiri*. 2020; 16(2) : 255–260.

[10]Adeniyi D A, Wei Z, Yongquan Y. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Appl. Comput. Informatics.* 2016; 12(1): 90–108.

[11] Nurjanah W E, Perdana R S, Fauzi M A. Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. *J. Pengemb. Technol. Inf. and Computer Science. Univ. Brawijaya*. 2017; 1(12): 1750–1757.

[12]Leidiyana H. Penerapan algoritma k-nearest neighbor untuk penentuan resiko kredit kepemilikan kendaraan bemotor. *PIKSEL (penelitian ilmu computer, system embedded & logic)* 2013; 1(1): 65–76.

[13] Lestari M. Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) untuk Mendeteksi Penyakit Jantung. *Fact. Exacta*. 2014; 7(4): 366–371.

[14] Hidayatul S, Aini A, Sari Y A, Arwan A. Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes. *JPTIIK (jurnal pengembangan teknologi infomarsi dan ilmu computer)* 2018; 2(9): 2546–2554.

[15] Yusra, Olivita D, Vitriani Y. Yusra, D. Olivita, and Y. Vitriani, “Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor. *SiTekIn (JURNAL SAINS, TEKNOLOGI DAN INDUSTRI UIN SYARIF KASIM RIAU)* 2016; 14(1): 79–85.

[16] Wella, Made N, Iswari S. Naive Bayes dalam Pengklasifikasian Kesegaran Ikan Menggunakan Media Foto. *ULTIMATICS (Jurnal Teknik Informatika)* 2017; 9(2): 114–117.

[17] Indriyanti, Sugianti D, Al Karomi M A. Peningkatan Akurasi Algoritma KNN dengan Seleksi Fitur G ain Ratio untuk Klasifikasi Penyakit Diabetes Mellitus. *IC-Tech*. 2017; 7(2): 1–6.

[18] Syafi'i S I, Wahyuningrum R T, Muntasa A. Segmentasi Obyek Pada Citra Digital Menggunakan Metode Otsu Thresholding. *J. Inform.* 2016; 13(1): 1–8.

[19] Ristoski P, Bizer C, Paulheim H. Mining the Web of Linked Data with Rapid Miner. *J. Web Semant.* 2015; 35(3): 142–151.

[20] Prasanti A A, Fauzi M A, Furqon M T. Klasifikasi Teks Pengaduan Pada Sambat Online Menggunakan Metode N- Gram dan Neighbor Weighted K-Nearest Neighbor (NW-KNN). *J. Pengemb. Technol. Inf. and Computer Science. Univ. Brawijaya*.2018; 2(2): 594–601.