Terbit online pada laman web jurnal: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

JURNAL DINDA

Kelompok Keahlian Rekayasa Data
Institut Teknologi Telkom Purwokerto

Vol. 1 No. 1 (2021) ISSN Media Elektronik: 12345-XYZ

Hasil Klasifikasi Algoritma *Backpropagation* dan *K-Nearest Neighbor* pada *Cardiovascular Disease*

Nashrulloh Khoiruzzaman[1], Rima Dias Ramadhani[2], Apri Junaidi[3]
[1]Program Studi S1 Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
[2,3]Program Studi S1 Sains Data, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
[1]16102097@ittelkom-pwt.ac.id, [2] rima@ittelkom-pwt.ac.id, [3]apri@ittelkom-pwt.ac.id

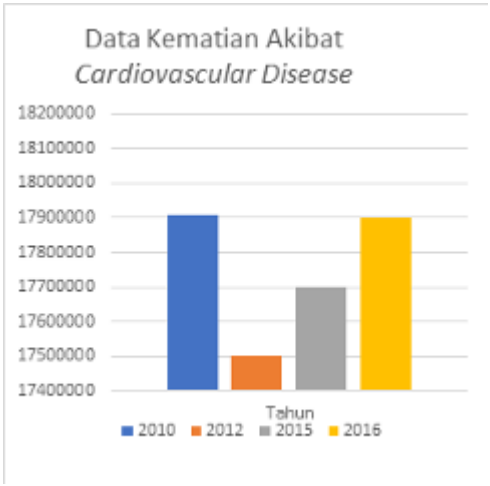
Abstract
Cardiovascular disease is a disease caused by abnormalities that occur in the heart organ, that can affect humans from young to old age, there are 13 factors that influence it, namely Age, Sex, Chest Pain, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, and Thal. Cardiovascular disease various types, including coronary heart disease, heart failure, high blood pressure, low blood pressure and others. Therefore, this study aims to classify cardiovascular disease. In this study using the backpropagation algorithm and the K-nearest neighbor algorithm. First step to do is the euclidean distance calculation process in K-NN to find the closest k distance to get the category based on the most frequent frequencies of the specified k value and look for new weights for the backpropagation algorithm to get new weights used to get values that are as expected. This system testing consists of testing the accuracy value with the K value, the K-fold X validation test and the hidden layer effect. The results of this study that the backpropagation algorithm produces an accuracy value of 64%, a precision of 62%, a recall of 64% and a K-nearest neighbor algorithm produce an accuracy value of 66%, a precision of 61% and a recall of 66%. The effect of hidden layer on the backpropagation algorithm in classifying cardiovascular disease is very large according to the results of research that have been conducted that when the number of hidden layers is small, the resulting value is also small but when the number of hidden layers is high the accuracy value is even low.
Keywords: backpropagation algorithm, k-nearest neighbor algorithm, cardiovascular disease, confusion matrix, neural network

Abstrak
Cardiovascular disease adalah penyakit yang diakibatkan oleh kelainan yang terjadi pada organ jantung, yang dapat menyerang manusia dari usia muda hingga usia tua yang terdapat 13 faktor yang mempengaruhinya yaitu Age, Sex, Chest pain, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, dan Thal. Cardiovascular disease beragam jenisnya antara lain penyakit jantung koroner, gagal jantung, tekanan darah tinggi, tekanan darah rendah dan lain-lain. Oleh karena itu, penelitian ini memiliki tujuan untuk melakukan klasifikasi terhadap cardiovascular disease. Pada penelitian ini menggunakan algoritma backpropagation dan algoritma K-nearest neighbor. Langkah awal dilakukan adalah proses perhitungan euclidean distance pada K-NN untuk mencari jarak k terdekat untuk mendapatkan kategori berdasarkan frequensi terbanyak dari nilai k yang ditentukan dan mencari bobot baru untuk algoritma backpropagation untuk mendapatkan bobot baru yang digunakan untuk mendapatkan nilai yang sesuai dengan yang diharapkan. Pengujian sistem ini terdiri dari pengujian nilai akurasi dengan nilai K, pengujian K-fold X validation dan pengaruh hidden layer. Hasil dari Penelitian ini bahwa algoritma backpropagation menghasilkan nilai akurasi sebesar 64%, presisi sebesar 62%, recall sebesar 64% dan algoritma K-nearest neighbor menghasilkan nilai akurasi sebesar 66%, presisi sebesar 61% dan recall sebesar 66%. Pengaruh hidden layer terhadap algoritma backpropagation dalam mengklasifikasikan cardiovascular disease sangat besar hal ini sesuai dengan hasil dari penelitian yang telah dilakukan bahwa ketika jumlah hidden layer kecil, nilai yang dihasilkan juga kecil akan tetapi ketika jumlah hidden layernya tinggi nilai akurasinya bahkan menjadi rendah..
Kata kunci: algoritma backpropagation, algoritma k-nearest neighbor, cardiovascular disease, confusion matrix, neural network

Diterima Redaksi : 13-12-2020 | Selesai Revisi : 21-12-2020 | Diterbitkan Online : 2-2-2021

1. Pendahuluan

Penelitian ini menganalisis hasil akurasi dari algoritma K-*nearest neighbor* dan *backpropagation* serta menganalisis pengaruh dari *hidden layer* terhadap klasifikasi *cardiovascular disease* dengan menggunakan dataset yang diperoleh dari website *UCI Machine Learning* pada alamat website https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease. *Cardiovascular disease* yang sering dijumpai yaitu penyakit jantung koroner. Hal ini, berdasarkan pada Survei *Sample Registration System* (SRS) yang dilakukan pada tahun 2014 yaitu sebesar 12,9% sedangkan berdasarkan data Riset Kesehatan Dasar (RISKESDAS) 2013 menunjukkan bahwa prevalensi tertinggi pada *cardiovascular disease* di Indonesia adalah penyakit jantung koroner sebesar 1,5 %. Berdasarkan data yang ada pada *World Health Organization* (WHO) pada tahun 2012 menunjukkan bahwa sebanyak 17,5 juta orang di dunia meninggal akibat *cardiovascular disease*[1].



Gambar 1. Data Kematian Akibat *Cardiovascular Disease*[1][2]

Jantung termasuk ke dalam golongan organ yang penting pada sistem tubuh manusia, karena jantung memiliki fungsi untuk memompa darah yang mengandung oksigen dan nutrien dari jantung ke seluruh tubuh dan kembali lagi ke jantung[3].

Jantung yang tidak berfungsi sebagaimana mestinya akan menimbulkan *cardiovascular disease* yang sangat banyak jenisnya antara lain penyakit jantung koroner, gagal jantung, tekanan darah tinggi, tekanan darah rendah dan penyakit jantung yang sangat berisiko dan sering menyerang orang dewasa antara lain penyakit jantung koroner dan penyakit gagal jantung[4].

Kemajuan teknologi pada ilmu komputer telah berkontribusi pada berbagai bidang, salah satunya adalah bidang kedokteran spesialis jantung dalam mengetahui jenis *cardiovascular disease* yang menyerang pasien. Penerapan teknologi tersebut diimplementasikan dalam metode klasifikasi menggunakan data mining. Data mining merupakan disiplin ilmu yang mempelajari metode pengolahan data yang dimaksudkan untuk menemukan pola yang tersembunyi pada data tersebut[5]. Data mining dikelompokkan menjadi 5 (lima) kelompok yaitu, *estimation, prediction, classification, clustering*, dan *asosiation* [6].

Classification merupakan proses penemuan model yang dapat mendeskripsikan dan membedakan kelas data supaya dapat digunakan untuk memprediksi kelas dari obyek yang label kelasnya tidak diketahui, klasifikasi ini terdiri dari 2 (dua) langkah yaitu, *learning* dan klasifikasi. Proses *learning* menggunakan algoritma klasifikasi dalam menganalisa data *training* yang direpresentasikan dalam bentuk *rule* klasifikasi. Proses klasifikasi menggunakan data uji dalam memperkirakan akurasi dari *rule* klasifikasi tersebut[6]. Klasifikasi terdapat beberapa algoritma antara lain, algoritma *naïve bayes*, algoritma *decision tree*, algoritma K-*Nearest Neighbor (K-NN)*, *logistic regresion*, dan *neural network*[7].

Cardiovascular disease akan mudah diketahui dengan melakukan prediksi secara dini dengan menggunakan sistem cerdas untuk membantu menekan angka kematian yang tinggi karena *cardiovascular disease*. Pada penelitian sebelumnya yang menggunakan algoritma K-NN mendapatkan nilai akurasi yang cukup tinggi[8]. Pada kasus lain penggunaan algoritma *backpropagation* juga menghasilkan nilai akurasi yang tinggi[9].

Penelitian yang dilakukan oleh M. Lestari [8]. Masalah dari penelitian ini adalah angka kematian yang tinggi berdasarkan data dari WHO, yaitu kematian yang diakibatkan oleh penyakit jantung. Untuk itu, diperlukan suatu pendeteksi dini penyakit jantung yang efektif dan akurat sebagai upaya mengatasi angka kematian yang tinggi akibat penyakit jantung. Pada penelitian ini menggunakan algoritma K-*nearest neighbor*, dan menggunakan data yang diperoleh dari University of California Irvine (UCI) *Machine Learning* data *repository*. Dari data tersebut diperoleh 14 atribut yang dapat digunakan dalam mendiagnosa penyakit jantung Hasil penelitian ini diperoleh nilai akurasi sebesar 70% serta nilai AUC sebesar 0,875 yang masuk kedalam klasifikasi baik, sehingga algoritma *K-Nearest Neighbor* dapat

digunakan dan diterapkan untuk mendeteksi penyakit jantung.

Penelitian yang dilakukan oleh Hidayatul S. dkk [10]. Menggunakan algoritma klasifikasi k-*nearest neighbor* dan algoritma *naïve bayes* dengan masalah tingginya angka kematian yang diakibatkan penyakit jantung, diperkirakan oleh Kementrian Kesehatan Republik Indonesia pada tahun 2030 hingga mencapai 23,3 juta penduduk. Ditambah lagi jumlah dokter penyakit jantung di Indonesia masih sangat sedikit. Untuk itu, diperlukan suatu sistem yang mampu membantu dokter yang kurang berpengalaman untuk mengetahui penyakit pada pasien. Ketika mengidentifikasi penyakit biasanya pasien mengikuti beberapa tes tetapi hasil dari tes tersebut tidak semua berkontribusi dengan diagnosis yang efektif, fitur yang tidak relevan dan berlebihan mengakibatkan hasil yang tidak akurat. Hasil penelitian ini menunjukkan nilai akurasi sebesar 92,31% pada saat pengujian sebaran kelas seimbang menggunakan 6 fitur dengan nilai K=25 dan pada saat pengujian sebaran kelas tidak seimbang menggunakan 4 fitur dengan nilai K=35. Fitur-fitur tersebut dipilih menggunakan metode *information gain* yang dapat memilih fitur yang paling sederhana dan dapat mengurangi noise.

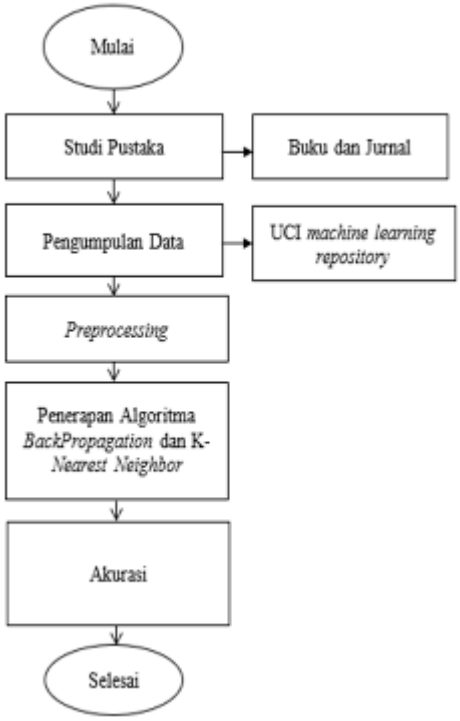
Penelitian yang dilakukan oleh Nawawi., H., M., dkk [9]. Menggunakan algoritma *neural network* dan algoritma *naïve bayes* dengan masalah penyakit jantung termasuk penyakit yang mematikan yang harus ditangani segera karena dapat terjadi secara tiba-tiba terhadap penderitanya. Oleh karena itu, penyakit jantung harus dideteksi sedini mungkin dengan mengacu kepada faktor-faktor yang dapat menyebabkan penyakit jantung. Hasil dari penelitian ini adalah algoritma *neural network* menghasilkan nilai akurasi sebesar 84,52% dengan rata-rata terkecil adalah 84,49% sedangkan algoritma *naïve bayes* dengan ditambah optimasi menghasilkan nilai akurasi sebesar 79,88% dengan nilai rata-rata terkecil sebesar 79,87%.

Penelitian yang dilakukan oleh Bachtiar Rifai [11]. Menggunakan algoritma *backpropagation* dalam melakukan prediksi awal penyakit jantung. Pada suatu industri kesehatan dan medis sangat membutuhkan keakuratan prediksi dari sebuah penyakit dan dengan keefektifan dari sebuah keputusan dari hasil analisa penyakit yang diderita pasien. Berdasarkan data yang ada pada penelitian ini, jumlah kematian yang disebabkan oleh penyakit jantung mencapai 959.227 pasien, sama dengan 41,4% dari seluruh kematian. Hasil dari penelitian ini mendapatkan nilai akurasi 91,45%, presisi 92,79%, *recall* 94,27%, dan nilai AUC 0,937.

Penelitian ini dilakukan untuk membuat komparasi hasil analisis dari klasifikasi *cardiovascular disease* untuk menentukan algoritma yang baik untuk membuat suatu sistem cerdas untuk mengurangi resiko kematian yang diakibatkan oleh *cardiovascular disease.* Sehingga dapat dilakukan pendeteksian secara dini. Pada penelitian ini dilakukan untuk memperoleh seberapa besar nilai akurasi dari algoritma *K-nearest neighbor* dan algoritma *backpropagation* kemudian mencari seberapa besar pengaruh *hidden layer* terhadap algoritma *backpropagation* dalam melakukan klasifikasi *cardiovascular disease.*

2. Metode Penelitian

Metode penelitian ini dilakukan dengan tahapan-tahapan dalam mengklasifikasikan *cardiovascular disease* dapat dilihat pada gambar 2 dibawah ini



Gambar 2 Flowchart Penelitian

2.1. Studi Pustaka

Pada tahap studi pustaka ini dilakukan pengumpulan sumber bacaan/ literatur dari penelitian sebelumnya dengan penggunaan algoritma yang sama, dengan obyek penelitian yang sama maupun berbeda. Sumber bacaan tersebut diperoleh dari beberapa jurnal ilmiah yang diperoleh dari google *scholar* (google cendekia) dan juga buku penunjang untuk algoritma yang digunakan yang diperoleh dengan membeli di *online shop.*

2.2. Pengumpulan data

Tahapan pengumpulan data ini menggunakan dataset yang diperoleh dari *website* UCI *Machine Learning*

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/, dataset yang diambil merupakan dataset *heart disease* yang ada pada *cleveland* dataset yang mempunyai 14 atribut. Atribut-atribut yang digunakan antara lain, umur, jenis kelamin, jenis nyeri dada, tekanan darah, kolesterol, kadar gula darah, hasil elektrokardiografi, denyut jantung maksimum yang tercapai, latihan yang diinduksi angina, depresi ST, kemiringan segmen latihan puncak ST, angka pada pewarnaan *flourosopy*, *thal real* dan diagnosis penyakit jantung. Dataset yang digunakan peneliti pada penelitian ini terdapat pada halaman web. https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Tabel 1 Atribut dan Skala Pengukuran

No.	Atribut	Skala Pengukuran
a.	*Age* (Usia)	[29,0;77,0]
b.	*Sex* (Jenis Kelamin)	[0,0;1,0] 0,0 = wanita 1,0 =pria
c.	*Chest Pain real*(Nyeri dada)	[1,0;4,0] 1 = Tipikal angina 2 = Angina atipikal 3 = Nyeri non angina 4 = Tanpa gejala
d.	*Trestbps(Resting blood pressure*/Tekanan Darah)	[94,0;200,0]
e.	*Chol*(serum kolestoral)	[126,0;564,0]
f.	*Fbs (fasting blood sugar)*/Gula Darah	[0,0;1,0] 0 = *false* 1 = *true*
g.	*Restecg* (hasil electrocardiografi)	[0,0;2,0] - Nilai 0 = normal. - Nilai 1 = memiliki kelainan gelombang ST-T (inversi gelombang T dan / atau elevasi atau depresi ST> 0,05 mV). - Nilai 2 = menunjukkan hipertrofi ventrikel kiri yang mungkin atau pasti berdasarkan kriteria Estes.
h.	*Thalach* (denyut jantung maksimum tercapai)	[71,0;202,0]
i.	*Exang* (latihan yang diinduksi angina)	[0,0;1,0] 0 = No 1 = Yes
j.	*Oldpeak* (Depresi ST disebabkan oleh olahraga relatif terhadap istirahat)	[0,0;6,2]
k.	*Slope* (kemiringan segmen latihan puncak ST)	[1,0;3,0] - Nilai 1 = Menanjak - Nilai 2 = Datar - Nilai 3 = Downsloping
l.	*Ca real* (*number of major vessels (0-3) colored by flourosopy*)	[0,0;3,0]
m.	*Thal real*	[3,0;7,0] 3 = Normal; 6 = Fixed defect; 7 = Reversable defect
n.	*Num* (diagnosis penyakit jantung)	{0;1;2;3;4} -Value 0: healthy -Value 1: low -Value 2: middle -Value 3: high -Value 4: seriously

2.3 Preprocessing data

Preprocessing data mempunyai beberapa cara antara lain, me*reaplace missing value* dan normalisasi [12]. Pada penelitian ini akan menggunakan teknik me*reaplace missing value* yang ada pada dataset *heart disease* yang diperoleh dari UCI *machine learning* secara online pada website https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. Teknik me*reaplace missing value* dapat dilakukan dengan mencari nilai tengah pada dataset tersebut kemudian mengganti data yang kosong dengan nilai tengah yang diperoleh dari persamaan berikut,

a. Untuk mencari nilai tengah dari data ganjil dapat dilakukan dengan persamaan dibawah ini,

$$\text{Me} = X\frac{n+1}{2} \qquad (1)$$

b. Untuk mencari nilai tengah dari data genap dapat menggunakan persamaan sebagai berikut,

$$\text{Me} = \frac{\left(\left(X\frac{n}{2}\right)+\left(X\frac{n}{2}+1\right)\right)}{2} \qquad (2)$$

Keterangan: Me = median (nilai tengah), X = variabel data, n = index data

Perhitungan untuk menghilangkan *missing value* pada kolom Ca terdapat 4 (empat) *missing value* pada data ke 156, 180, 263, dan 274 sebagai berikut,

Tabel 2 Perhitungan *Missing value* dari data atribut Ca

No.	155	156	...	304
Age	57,0	52,0	...	57,0
Sex	1,0	1,0	...	0,0
Cp	4,0	3,0	...	4,0
Trestbps	132,0	138,0	...	140,0

Chol	207,0	233,0	...	241,0
Fbs	0,0	0,0	...	0,0
Restecg	0,0	0,0	...	0,0
Thalach	168,0	169,0	...	123,0
Exang	1,0	0,0	...	1,0
Oldpeak	0,0,	0,0	...	0,2
Slope	1,0	1,0	...	2,0
Ca	0,0	?	...	0,0
Thal	7,0	3,0	...	7,0
Num	0	0	...	1

Tabel 2 merupakan tabel dataset dengan *missing value* pada baris *Ca* yang akan diproses dengan persamaan median (nilai tengah) untuk mengganti *missing value* tersebut dengan nilai median (nilai tengah) yang telah diperoleh sebagai berikut:

$$\text{Me} = \frac{((X\frac{n}{2})+(X\frac{n}{2}+1))}{2}$$
$$= \frac{(\left(X\frac{304}{2}\right)+\left(X\frac{304}{2}+1\right))}{2}$$
$$= \frac{((X\ 152)+(X\ 153))}{2}$$
$$= \frac{((0,0)+(0,0))}{2}$$

= 0,0

Tabel 3 Hasil Perhitungan *Missing value* dari data atribut Ca

No.	155	156	...	304
Age	57,0	52,0	...	57,0
Sex	1,0	1,0	...	0,0
Cp	4,0	3,0	...	4,0
Trestbps	132,0	138,0	...	140,0
Chol	207,0	233,0	...	241,0
Fbs	0,0	0,0	...	0,0
Restecg	0,0	0,0	...	0,0
Thalach	168,0	169,0	...	123,0
Exang	1,0	0,0	...	1,0
Oldpeak	0,0,	0,0	...	0,2
Slope	1,0	1,0	...	2,0
Ca	0,0	**0,0**	...	0,0
Thal	7,0	3,0	...	7,0
Num	0	0	...	1

Tabel 3 diatas merupakan hasil dari penggantian *missing value* pada salah satu *missing value* pada baris Ca dengan nilai median (nilai tengah) sedangkan perhitungan untuk *missing value* pada atribut *Thal* pada data ke 82 dan 246 dapat dilakukan dengan persamaan sebagai berikut:

Tabel 4 Perhitungan *Missing value* dari atribut Thal

No.	81	82	...	304
Age	47,0	53,0	...	57,0
Sex	1,0	0,0	...	0,0
Cp	3,0	3,0	...	4,0
Trestbps	138,0	128,0	...	140,0
Chol	257,0	216,0	...	241,0
Fbs	0,0	0,0	...	0,0
Restecg	2,0	2,0	...	0,0
Thalach	156,0	115,0	...	123,0
Exang	0,0	0,0	...	1,0
Oldpeak	0,0	0,0	...	0,2
Slope	1,0	1,0	...	2,0
Ca	0,0	0,0	...	0,0
Thal	3,0	?	...	7,0
Num	0	0	...	1

Tabel 4 merupakan tabel dataset dengan *missing value* pada baris *Thal* yang akan diproses dengan persamaan median (nilai tengah) untuk mengganti *missing value* tersebut dengan nilai median (nilai tengah) yang telah diperoleh sebagai berikut :

$$\text{Me} = \frac{((X\frac{n}{2})+(X\frac{n}{2}+1))}{2}$$
$$= \frac{(\left(X\frac{304}{2}\right)+\left(X\frac{304}{2}+1\right))}{2}$$
$$= \frac{((X\ 152)+(X\ 153))}{2}$$
$$= \frac{((3,0)+(3,0))}{2}$$

= 3,0

Tabel 5 Hasil Perhitungan *Missing value Pada atribut Thal*

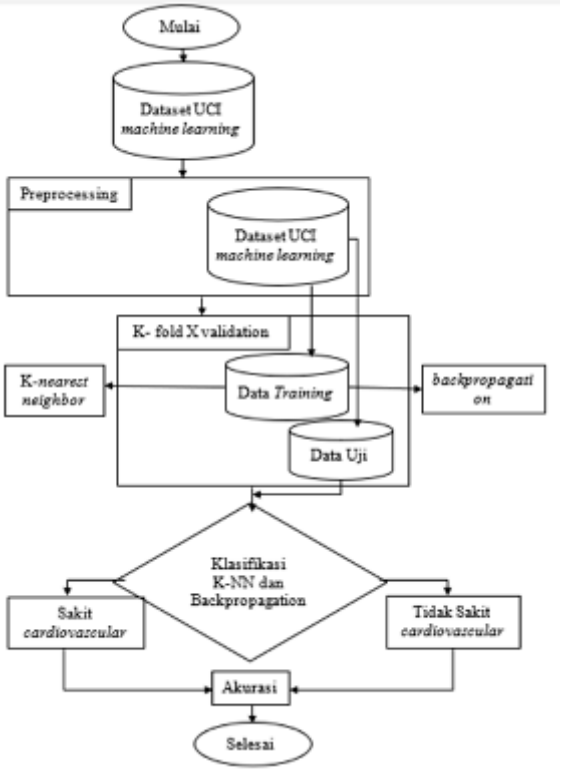
No.	81	82	...	304
Age	47,0	53,0	...	57,0
Sex	1,0	0,0	...	0,0
Cp	3,0	3,0	...	4,0
Trestbps	138,0	128,0	...	140,0
Chol	257,0	216,0	...	241,0
Fbs	0,0	0,0	...	0,0
Restecg	2,0	2,0	...	0,0
Thalach	156,0	115,0	...	123,0
Exang	0,0	0,0	...	1,0
Oldpeak	0,0	0,0	...	0,2
Slope	1,0	1,0	...	2,0
Ca	0,0	0,0	...	0,0
Thal	3,0	**3,0**	...	7,0
Num	0	0	...	1

Tabel 5 merupakan hasil penggantian salah satu *missing value* pada baris *Thal* dengan nilai median (nilai tengah) selanjutnya semua nilai yang bernilai <null> atau dengan tanda baca tanda tanya (?), diganti dengan nilai dari hasil perhitungan seperti yang ada di atas. Penjelasan diatas merupakan cara yang dilakukan secara manual, apabila dengan menggunakan pemrograman python dapat di tuliskan seperti pada program berikut ini:

Program Jurnal

```
Input: Dataset UCI
Output: Median
  ###Memilih kolom data yang ada missing
  valuenya
     X=data.iloc[:,11].values #Menunjukkan
     kolom dengan index 11 atau kolom Ca
     Y=data.iloc[:,12].values #Menunjukkan
     kolom dengan index 12 atau kolom Thal
     print("data Ca: \n",X)
     print("data Thal : \n",Y)
     #Menampilkan gambar letak missing value
     sns.heatmap(data.isnull(),cbar=False)
     plt.title('Heatmap Missing Value')
     plt.show()
     #Menampilkan Persentase dari Missing
     value tersebut
     (data.isnull().sum()/len(data)).to_frame(
     'persentase missing')
     #Menampilkan Mising Value pada Kolom Ca
     dan Thal
     #dan
     #Menampilkan Hasil Nilai Median
     data['Ca_imputed_median'] =
     data['Ca'].replace(np.nan,
     data.Ca.median())
     a=data[['Ca','Ca_imputed_median']].head(2
     74)
     data['Thal_imputed_median'] =
     data['Thal'].replace(np.nan,
     data.Thal.median())
     b=data[['Thal','Thal_imputed_median']].he
     ad(84)
     print("Replace data Ca: \n",a)
     print("Replace data Thal: \n",b)
     #Mengganti Missing Value dengan Nilai
     Median yang didapatkan
     median = data["Ca"].median()
     data["Ca"] = data["Ca"].replace(np.nan,
     median)
     median = data["Thal"].median()
     data["Thal"] =
     data["Thal"].replace(np.nan, median)
     ca=data.Ca.head(274)
     thal=data.Thal.head(83)
     print("Hasil replace data Ca: \n",ca)
     print("Hasil replace data Thal: \n",thal)
```

2.4 Penerapan Algoritma K-*Nearest Neighbor* dan algoritma *backpropagation*



Gambar 3 Flowchart Algoritma *K-Nearest Neighbor* dan algoritma *backpropagation*

Berdasarkan pada Gambar 3.2 diatas dapat dijelaskan bahwa dataset yang diperoleh dari UCI Machine Learning, diolah dengan proses yang disebut dengan preprocessing data. Preprocessing data dilakukan dengan tujuan untuk menghilangkan missing value dari dataset tersebut. Kemudian hasil dari preprocessing data, kemudian dilanjutkan ke tahap berikutnya yaitu melakukan pengujian K-fold x validation dengan menginputkan data training dan data uji, pengujian K-fold x validation tersebut dilakukan dengan menggunakan 2 (dua) algoritma yaitu algoritma K-nearest neighbor dan algoritma backpropagation. Selanjutnya dilakukan pengklasifikasian apakah pasien menderita cardiovascular disease atau tidak, kemudian dilakukan

pengujian tingkat akurasi untuk mendapatkan hasil dari penelitian ini.

3. Hasil dan Pembahasan

3.1. Pengumpulan data

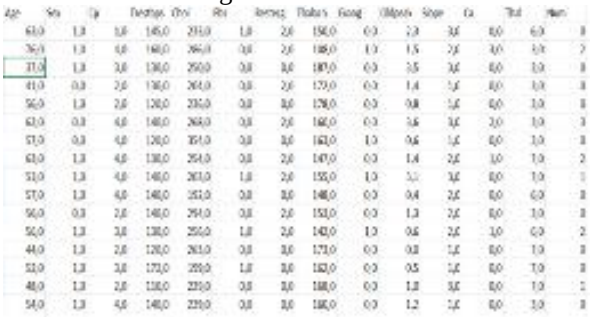
3.1.1 Pengunduhan data

Pada penelitian ini pengumpulan data dilakukan dengan mengunduh dataset yang telah tersedia pada website UCI *machine learning*. Dataset yang diambil merupakan data *heart disease*, data yang diperoleh dari website tersebut berjumlah 304 data dengan 6 data *missing value* yang terbagi pada 2 (dua) kategori yaitu *Ca* dan *Thal*. Dari hasil pengunduhan dataset *heart disease* tersebut berformat *file* .data, berikut ini merupakan beberapa contoh data *heart disease* yang diunduh pada website UCI *machine learning*:

Tabel 6. Dataset heart disease berformat .data

63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,0
67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,2
67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,1
37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,0
41.0,0.0,2.0,130.0,204.0,0.0,2.0,172.0,0.0,1.4,1.0,0.0,3.0,0
56.0,1.0,2.0,120.0,236.0,0.0,0.0,178.0,0.0,0.8,1.0,0.0,3.0,0
62.0,0.0,4.0,140.0,268.0,0.0,2.0,160.0,0.0,3.6,3.0,2.0,3.0,3
57.0,0.0,4.0,120.0,354.0,0.0,0.0,163.0,1.0,0.6,1.0,0.0,3.0,0
63.0,1.0,4.0,130.0,254.0,0.0,2.0,147.0,0.0,1.4,2.0,1.0,7.0,2
53.0,1.0,4.0,140.0,203.0,1.0,2.0,155.0,1.0,3.1,3.0,0.0,7.0,1

Pada data tersebut terdapat 13 (tiga belas) atribut yang dapat mempengaruhi *cardiovasculardisease* antara lain, Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal dan 1 (satu) atribut sebagai kategori yaitu Num. Untuk memahami dataset tersebut format data diubah ke dalam format dokumen excel sebagai berikut:



Gambar 4 Dataset heart disease berformat excel

3.2. preprocessing

Proses selajutnya yaitu *pre-processing*, proses ini dilakukan untuk memperbaiki data yang kosong atau bisa disebut *missing value*. Cara mengatasi *missing value* dapat dilakukan dengan cara me*reaplace* data. Teknik me*reaplace* data pada penelitian ini yaitu dengan menggunakan nilai tengah pada data atau median. Teknik me*reaplace* data dengan menggunakan nilai tengah bisa menggunakan program maupun manual, tetapi pada penelitian ini menggunakan nilai tengah yang dijalankan program dengan bahasa pemrograman *python.*

3.3. Klasifikasi

Pada tahapan ini dataset yang sudah dilakukan *pre-processing* akan masuk kedalam tahapan klasifikasi menggunakan algoritma K-*Nearest Neighbor* dan algoritma *Backpropagation*, tahapan ini dilakukan untuk mengetahui seberapa besar nilai akurasi yang diberikan oleh algoritma tersebut. Proses ini dilakukan dengan memisahkan data menjadi data training dan data test. Data training pada penelitian ini sebesar 80% dari jumlah dataset dan data test sebesar 20% dari dataset.

Berikut ini merupakan pembahasan dari penelitian yang telah dilakukan sebagai berikut:

a. Algoritma K-nearest neighbor

Pada penelitian ini dalam mendapatkan nilai akurasi dari algoritma k-*nearest neighbor* dengan menjalankan algoritma tersebut. Pada penelitian ini sesuai dengan batasan masalah yang ada yaitu dengan menggunakan bahasa pemrograman *python* dan dengan *tools* yang disediakan oleh google yaitu *google colaboratory.* Langkah awal yang dilakukan adalah mendeskripsikan dataset *cardiovascular disease*. Kemudian melakukan pelabelan pada target yang diuji, pada dataset ini terdapat 5 (empat) kategori yaitu dengan nilai 0-4. Nilai 0 menunjukan *Healthy*, nilai 1 menunjukan *low*, nilai 2 menunjukan *middle*, nilai 3 menunjukan *high* dan nilai 4 menunjukan *seriously*. Dataset *cardiovascular disease* diolah dengan menghitung jarak k terdekat menggunakan *euclidean distance.* Hasil perhitungan *euclidean distance* dapat dilihat pada Tabel 7:

Tabel 7. Hasil perhitungan euclidean distance

No.	Data Test	Data Train				
		1	2	3	4	5

Euclidean Distance	1	72,03950305	124,1225604	107,4570147	53,13981558	89,87124123
	2	88,81019086	53,70372427	90,24904432	120,77773158	97,56023 78

Hasil perhitungan euclidean distance tersebut dataset diurutkan dari jarak k minimul ke jarak k maksimum, yamng mana k pada penelitian ini adalah 10 (sepuluh). Kesepuluh data tersebut dicarikan kategori yang frequensinya terbanyak sehingga didapatkan hasil sebagai berikut:

Tabel 8. Hasil kategori dengan frequensi terbanyak

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
d	18,32	19,59	20,08	21,48	21,49	23,43	23,58	24,24	24,74	25,52
Kategori	0	0	0	0	1	0	0	0	0	0

Pada penelitian ini menghitung jarak dengan menggunakan rumus *euclidean distance* untuk mengukur jarak antara data training dengan data test sehingga menghasilkan akurasi dari algoritma K-*nearest neighbor* sebesar 66%, nilai presisinya sebesar 61% dan nilai recallnya sebesar 66%.

b. Algoritma Backpropagation

Algoritma *backpropagation* merupakan algoritma *neural network* atau yang disebut dengan multilayer perceptron yang mana model tersebut lebih sederhana dari *neural network* yang dapat digunakan untuk menyelesaikan tugas komputasi yang sulit seperti dalam *machine learning. Algoritma backpropagation mempunyai tahapan proses sebagai berikut:* Menentukan bobot awal secara acak, *learning rate* dan target *output.* Menghitung nilai output pada *hidden layer*. Menghitung *output* pada *output layer*, proses pertama sampai ketiga disebut proses *forward propagation.* Berikut ini merupakan hasil dari perhitungan dari *forward propagation.* Berikut ini merupakan hasil perhitungan dari algoritma *Backpropagation:*

Tabel 9. Hasil perhitungan algoritma *backpropagation*

Learning rate = 0.005, Target output = 1			
Input Layer	**Hidden Layer**	**Bobot Ke Output Acak**	**Output Layer**

Input	**1**	**Bobot Acak**	**Output Node 1**	**Output Node 2**		
Age	63,0	$W1_1$=0,1 $W1_2$= -0,02				
Sex	1,0	$W2_1$= -0,3 $W2_2$= -0,5				
Cp	1,0	$W3_1$= 0,3 $W3_2$= 0,1				
Trestbps	145,0	$W4_1$= 0,01 $W4_2$= 0,09				
Chol	233,0	$W5_1$= -0,06 $W5_2$= -0,05				
Fbs	1,0	$W6_1$= 0,7 $W6_2$= -0,6	0,545	0,50533	Wn_1 = -0,3 Wn_2 = 0,1	0,890134
Restecg	2,0	$W7_1$= 0,3 $W7_2$= -0,4				
Thalach	150,0	$W8_1$= -0,1 $W8_2$= -0,01				
Exang	0,0	$W9_1$= -0,2 $W9_2$= -0,1				
Oldpeak	2,3	$W10_1$= 0,5 $W10_2$= -0,2				
Slope	3,0	$W11_1$= 0,5 $W11_2$= -0,2				
Ca	0,0	$W12_1$=				

		0,8 $W12_2$= 0,7				
Thal	6,0	$W13_1$= 0,6 $W13_2$= 0,7				

Proses tersebut dilanjutkan dengan menghitung nilai *error* pada *output layer* dan *hidden layer*. Berikut ini merupakan hasil dari perhitungan nilai *error output layer* dan *hidden layer:*

Tabel 10. Hasil perhitungan nilai error pada *output layer* dan *hidden layer*

Hidden Layer		**Bobot Ke Output Acak**	**Output Layer**	**Nilai Error dari Hidden Layer Ke Output Layer**		
Output Node 1	**Output Node 2**			**Error Output**	**Error Node 2**	**Error Node 1**
0,545	0,50533	Wn_1 = -0,3 Wn_2 = 0,1	0,890134	0,0865621	-0,12506835	-0,1339065

Menghitung bobot baru dari output layer ke hidden layer, bobot dari output ke hidden layer digunakan untuk melanjutkan proses backpropagation dan menggunakan bobot baru tersebut sebagai iterasi selanjutnya. Berikut ini merupakan hasil perhitungan dari bobot baru yang telah diperoleh:

Tabel 11. Bobot baru yang dihasilkan untuk melanjutkan iterasi

Bobot Baru	
Bobot dari Output Ke Hidden Layer	**Bobot Untuk Melanjutkan Iterasi**
Wn_1 = -0,299 Wn_2 = 0,100	$W1_1$= 0,058 $W1_2$= -0,059
	$W2_1$= -0,300 $W2_2$= -0,500
	$W3_1$= 0,299 $W3_2$= 0,099
	$W4_1$= -0,087 $W4_2$= -0,006
	$W5_1$= -0,216 $W5_2$= -0,195
	$W6_1$= 0,699 $W6_2$= -0,600
	$W7_1$= 0,298 $W7_2$= -0,401
	$W8_1$= 0,200 $W8_2$= -0,104
	$W9_1$= -0,2 $W9_2$= -0,1
	$W10_1$= 0,498 $W10_2$= -0,201
	$W11_1$= 0,497 $W11_2$= -0,201
	$W12_1$ = 0,8 $W12_2$= 0,7
	$W13_1$= 0,595 $W13_2$= 0,696

Pada algoritma *backpropagation* ini menggunakan teknik multilayer perceptron dimana teknik tersebut menggunakan sistem kerja *backpropagation* dengan fungsi sigmoid di wakilkan dengan *activation logistic*, *hidden layer*nya berjumlah 1024 dan *learning rate*nya sebesar 0.005 sehingga menghasilkan nilai akurasi sebesar 64%, presisinya sebesar 62%, dan recallnya sebesar 64%.

c. Akurasi

Pada penelitian ini dalam menentukan nilai akurasi dengan menggunakan pengujian K *fold X validation*, *K-fold X validation* merupakan metode yang biasa digunakan dalam mengetahui rata-rata keberhasilan dari suatu sistem. Sistem melakukan perulangan dengan mengacak *input* sehingga sistem tersebut teruji untuk beberapa *input* yang acak. Metode K-*fold X validation* melakukan pembagian sampel secara aktual sehingga k menjadi sampel yang berukuran sama. Data validasi diperoleh dari subsampel yang digunakan dalam melakukan pengujian model klasifikasi dengan mengulangi proses tersebut sebanyak k kali. Pada penelitian ini menggunakan jumlah *fold* sebanyak (10) sepuluh *fold*, maka dari itu dataset dibagi menjadi (10) sepuluh bagian, berikut ini merupakan cara kerja sepuluh *fold* X *validation*:

Tabel 12. Cara Kerja 10 *fold X validation*

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Dari hasil evaluasi tersebut didapatkan nilai akurasi pada algoritma *backpropagation* sebesar 64% dan K-*Nearest Neighbor* sebesar 67% pada saat k = 11. Berikut ini merupakan hasil akurasi dari K-*nearest neighbor* untuk nilai k= 1-20:

Tabel 13. Nilai k beserta nilai akurasi

K	Akurasi (%)
1	54
2	54
3	57
4	62
5	66
6	64
7	60
8	64
9	64
10	66
11	67
12	64
13	64
14	62
15	62
16	62
17	60
18	64
19	64
20	64
21	64
22	64
23	64

24	64
25	64

Penelitian yang telah dilakukan menghasilkan akurasi dari algoritma K-*nearest neighbor* sebesar 66%, nilai presisinya sebesar 61% dan nilai recallnya sebesar 66%. Pada algoritma *backpropagation* ini menggunakan teknik multilayer perceptron dimana teknik tersebut menggunakan sistem kerja *backpropagation* dengan fungsi sigmoid di wakilkan dengan *activation logistic*, *hidden layer*nya berjumlah 1024 dan *learning rate*nya sebesar 0.005 sehingga menghasilkan nilai akurasi sebesar 64%, presisinya sebesar 62%, dan recallnya sebesar 64%. Dari hasil penelitian tersebut dapat diketahui bahwa algoritma K-*nearest neighbor* mempunyai nilai akurasi lebih tinggi dari algoritma *backpropagation* dengan selisih 2%.

d. Pengaruh *hidden layer* pada *backpropagation*
Pada penelitian ini dalam menentukan hidden layer dilakukan secara acak dengan menggunakan jumlah hidden layer mulai dari angka puluhan hingga ribuan secara manual sehingga ditemukan nilai hidden layer yang menghasilkan nilai akurasi yang tinggi pada nilai 1024 sehingga dapat diketahui bahwa *hidden layer* pada *backpropagation* sangat berpengaruh. Hal ini sesuai dengan penelitian bahwa ketika kondisi *hidden layer* rendah nilai akurasi yang dihasilkan rendah begitu juga pada saat *hidden layer*nya lebih tinggi nilai akurasi yang dihasilkan rendah juga. Dengan demikian, artinya *hidden layer* berpengaruh terhadap nilai akurasi yang dihasilkan pada proses pengklasifikasian.

4. Kesimpulan

Pada penelitian yang telah dilakukan dapat diperoleh kesimpulan sebagai berikut:
Algoritma K-*nearest neighbor* mendapatkan nilai akurasi sebesar 69% sedangkan algoritma *backpropagation* mendapatkan nilai akurasi sebesar 64% dari data training 80% dan data uji sebesar 20% sehingga dapat disimpulkan bahwa pada penelitian ini algoritma K-*nearest neighbor* lebih baik daripada algoritma *backpropagation* dalam mengklasifikasikan *cardiovascular disease*.
Pengaruh *hidden layer* pada algoritma *backpropagation* sangat tinggi tetapi tidak dapat dijelaskan secara matematis, hanya saja dapat dijelaskan dengan pernyataan bahwa semakin rendah jumlah *hidden layer* nilai akurasi yang dihasilkan juga rendah dan ketika jumlah *hidden layer*nya tinggi nilai akurasi yang dihasilkan juga rendah sehingga jumlah *hidden layer* harus ditentukan secara acak dan manual.

Ucapan Terimakasih
Berikut ini merupakan ucapan terima kasih atas bantuan dana dan fasilitas yang telah diberikan kepada penulis, ucapan terima kasih ditujukan kepada :
1. Bapak Waluyo Suprianto Sebagai Bapak dari peneliti.
2. Ibu Robiyah Sebagai Ibu dari peneliti.

Daftar Rujukan

[1]KEMENTRIAN KESEHATAN RI, “Penyakit Jantung Penyebab Kematian Tertinggi,” *Kementrian Kesehat. Republik Indones.*, pp. 1–2, 2017.
[2]W. H. Organization, *World Health Statistics 2018*. 2018.
[3]D. Ramli and Y. Karan, “Anatomi dan Fisiologi Kompleks Mitral,” *J. Kesehat. Andalas*, vol. 2, no. 7, pp. 103–112, 2018.
[4]RISKESDAS, “Riskesdas 2013,” *Jakarta Badan Penelit. dan Pengemb. Kesehat. Dep. Kesehat. Republik Indones.*, pp. i–268, 2013
[5] S. Heni and A. Irham Gufroni, “Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 03, no. 02, pp. 299–305, 2017.
[6]A. Maulana and A. A. Fajrin, “Penerapan Data Mining Untuk Analisis Pola Pembelian Konsumen Dengan Algoritma Fp-Growth Pada Data Transaksi Penjualan Spare Part Motor,” *Klik - Kumpul. J. Ilmu Komput.*, vol. 5, no. 1, p. 27, 2018
[7]S. Dewi, “Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan,” *Techno Nusa Mandiri*, vol. XIII, no. 1, pp. 60–66, 2016
[8]M. Lestari, “Penerapan Algoritma Klasifikasi Nearest Neighbor (K-Nn) Untuk Mendeteksi Penyakit Jantung,” *Fakt. Exacta*, vol. 7, no. 4, pp. 366–371, 2014.
[9]H. M. Nawawi *et al.*, “Komparasi Algoritma Neural Network dan Naive Bayes untuk Memprediksi Penyakit Jantung,” *J. PILAR Nusa Mandiri*, vol. 15, no. 2, pp. 189–194, 2019.
[10]S. Hidayatul, A. Aini, Y. A. Sari, and A. Arwan, “Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 2546–2554, 2018 .
[11]B. Rifai, “Algoritma Neural Network Untuk Prediksi Penyakit Jantung,” *Techno Nusa Mandiri*, vol. IX, no. 1, pp. 1–9, 2013.
[12]A. Fagustina, Y. Palgunadi, and Wiharto, “Pengaruh Fungsi Pembelajaran Terhadap Kinerja Pelatihan Jaringan Syaraf Tiruan Backpropagation,” no. March, 2018.