p-ISSN 2085-7829
e-ISSN 2798-3455

Klasifikasi Status Hipertensi Pasien UPTD Puskesmas Sempaja, Kota Samarinda Menggunakan Metode *K-Nearest Neighbor*

Classification of Hypertension Status of Patients UPTD Health Center Sempaja, Samarinda City Using the K-Nearest Neighbor Method

Raihana Soraya Putri Pratama[1, a)], Memi Nor Hayati[1, b)], dan Rito Goejantoro[2, c)]
[1]Laboratorium Statistika Ekonomi dan Bisnis FMIPA Universitas Mulawarman
[2]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman

[a)]raihanasoraya09@gmail.com
[a)]*Corresponding author*: meminorhayati@fmipa.unmul.ac.id
[c)]rito.goejantoro@fmipa.unmul.ac.id

ABSTRACT

Data mining is a method of selecting, exploring and modeling large amount of data to find knowledge and clear patterns or interesting relation of the data and useful in the process of data analysis. In data mining there are several techniques that have different function and one of them is classification tehcnique. The classification process itself is the process of finding patterns or differences between classes or data that can be used to predict object classes whose class labels are unknown. K-nearest neighbor (K-NN) is one of the methods in classification algorithm. This study discusses the classification using K-NN algorithm which is applied to the data hypertension status. The aim is to find out the optimal neighborliness value (K) accuracy value and the best propotion of the data hypertension status. The data used is the data of patients UPTD health center Sempaja, Samarinda city from February to May 2022 with dependent variabel is hypertension status and uses 4 independent variables, age, gender, diabetes mellitus and heart disease. Based on the research that has been done, obtained an accuracy value of 62,60% with K = 5 in the best proportion of the data is 70%:30%.

Keywords: *Classification, Data Mining, K-Nearest Neighbor*

1. Pendahuluan

Kesehatan adalah situasi yang diinginkan dari semua orang, di mana keadaan yang sejahtera pada tubuh, jiwa dan sosial memungkinkan semua orang untuk hidup produktif dalam bentuk sosial maupun secara ekonomi (Tjiptoherijanto, 1994). Tekanan darah tinggi atau hipertensi adalah salah satu masalah kesehatan yang dihadapi oleh masyarakat di negara-negara maju dan berkembang (Sidabuntar, 2009). Hipertensi merupakan penyakit tidak menular yang menjadi salah satu risiko kematian tertinggi ke-3 di Indonesia yakni sebesar 23,7% dari 1,7 juta kematian di Indonesia. (Yogiantoro, 2009). Hipertensi dapat terjadi dikarenakan faktor tertentu, seperti usia, kebiasaan merokok, stres, berat badan berlebih, konsumsi alkohol, hingga kurangnya aktivitas fisik (Nyoman, 2016).

Data mining adalah metode pemilihan, eksplorasi dan pemodelan data dalam jumlah besar untuk menemukan pengetahuan dan pola atau hubungan yang jelas, menarik dan bermanfaat dalam proses analisis data (Han dkk, 2012). Dalam data mining terdapat beberapa teknik yang memiliki fungsi berbeda dan salah satunya adalah teknik klasifikasi (Annur, 2018).

Klasifikasi adalah suatu bentuk analisis data yang dapat menggambarkan pola yang diekstraksi dari sebagian besar data. (Han dkk, 2012). Berdasarkan metode pelatihan, algoritma klasifikasi dapat dibedakan menjadi dua jenis, yaitu *eager learner* dan *lazy learner*. Algoritma klasifikasi yang termasuk dalam kategori *lazy learner* yaitu Regresi Linear, *K-Nearest Neighbor*, *Fuzzy K-Nearest Neighbor*, dan sebagainya (Prasetyo, 2014).

Algoritma *K-Nearest Neighbor* atau K-NN bertujuan untuk mengklasifikasikan subjek baru berdasarkan data *training* sampel dan variabel. Setelah itu, hasil sampel baru diklasifikasikan menurut mayoritas dari kelompok pada K-NN. Algoritma K-NN menggunakan klasifikasi yang bergantung pada data tetangganya yang digunakan sebagai nilai prediktor sampel uji baru (Krisandi dkk, 2013).

Dilakukan penelitian menggunakan variabel dependen adalah status hipertensi pasien dengan variabel bebas yaitu usia pasien, jenis kelamin pasien, status penyakit jantung pasien dan status diabetes mellitus pasien dengan nilai K = 3, 5, 7 dan 9 serta proporsi data *training* dengan data *testing* yaitu 90:10, 70:30 dan 50:50. Adapun tujuan digunakan beberapa nilai K dan beberapa proporsi data *training* dengan data *testing* ialah untuk memperoleh hasil klasifikasi berdasarkan metode K-NN dengan nilai K optimum dan proporsi.

p-ISSN 2085-7829
e-ISSN 2798-3455

2. Tinjauan Pustaka

2.1 Data Mining

Data mining adalah proses pemilihan, mengekplorasi dan pemodelan data dalam jumlah besar untuk menemukan pola atau hubungan yang jelas dan berguna dalam proses analisis data (Bellazzi & Zupan, 2008). *Data mining* memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan *database*. Beberapa metode yang sering disebut-sebut dalam literatur *data mining* antara lain *clustering*, *classification*, *association rules mining*, *neural network*, *genetic algorithm* dan lain-lain (Han dkk, 2012).

2.2 Klasifikasi

Klasifikasi merupakan metode yang digunakan untuk menemukan model atau fungsi yang digambarkan dengan perbedaan kelas data atau konsep yang berfungsi untuk memprediksi kelas dari objek yang label sudah diketahui. Proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Annur, 2018).

Model klasifikasi kemudian dibangun berdasarkan data *training* dan kemudian kinerja diukur berdasarkan data *testing*. Proporsi adalah bagian (persentase) suatu kejadian khususnya dari keseluruhan data yang ada. Perhitungan menggunakan persamaan sebagai berikut :

$$tr = p \times n \tag{1}$$

$$te = n - tr \tag{2}$$

Keterangan :
tr = Banyaknya data *training*
te = Banyaknya data *testing*
p = Proporsi data *training*
n = Banyaknya data pengamatan

2.3 K-Nearest Neighbor

K-nearest Neighbor atau K-NN merupakan salah satu metode berbasis klasifikasi Nearest Neighbor yang paling tua dan populer. Nilai K yang digunakan disini menyatakan jumlah tetangga terdekat yang dilibatkan dalam penetuan prediksi label kelas pada data uji. (Prasetyo, 2014).

Algoritma K-NN memiliki beberapa kelebihan yaitu ketangguhan terhadap data *training* yang memiliki banyak *noise* dan efektif apabila *training* datanya besar. Sedangkan, kelemahan K-NN adalah K-NN perlu menentukan nilai dari parameter K (banyaknya tetangga terdekat), *training* berdasarkan jarak tidak jelas mengenai jenis jarak apa yang harus digunakan dan variabel mana yang harus digunakan untuk mendapatkan hasil terbaik, dan biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap *query instance* pada keseluruhan sampel data *training* (Yustanti, 2012).

Perhitungan jarak untuk menentukan tetangga terdekat dapat menggunakan jarak Gower. Menurut Prasath dkk (2017), jarak Gower adalah jarak yang ditransformasi dari Koefisien kemiripan Gower. Koefisien kemiripan Gower merupakan koefisien kemiripan yang dikemukakan oleh Gower pada tahun 1971. Koefisien kemiripan Gower dapat digunakan untuk melakukan perhitungan kemiripan pada setiap variabel yang ada sesuai dengan skala pengukuran variabel tersebut. Perhitungan jarak Gower pada skala data nominal, ordinal, interval, dan rasio dapat dihitung dengan persamaan berikut.

- Jarak Gower pada data skala nominal,

$$d_q(a,b) = \begin{cases} 0, & x_{aq} = x_{bq} \\ 1, & x_{aq} \neq x_{bq} \end{cases} \tag{3}$$

- Jarak Gower pada data skala ordinal,

$$d_q(a,b) = \frac{\left|R_q(a) - R_q(b)\right|}{\text{maks}\left(R_q\right) - \min\left(R_q\right)} \tag{4}$$

- Jarak Gower pada data skala interval dan rasio,

$$d_q(a,b) = \frac{\left|x_{aq} - x_{bq}\right|}{\text{maks}\left(x_q\right) - \min\left(x_q\right)} \tag{5}$$

Maka, perhitungan jarak antara kedua objek tersebut,

$$d(a,b) = \frac{\sum_{q=1}^{r} d_q(a,b)}{r} \tag{6}$$

Keterangan :
$d_{q(a,b)}$ = Jarak antara data *training* ke-a dengan data *testing* ke-b pada variabel ke-q dengan $q = 1,2,\ldots,r$
x_{aq} = Nilai data *training* ke-a pada variabel ke-q

p-ISSN 2085-7829
e-ISSN 2798-3455

x_{bq} = Nilai data *testing* ke-*b* pada variabel ke-*q*
r = Banyaknya variabel bebas
$R_q(a)$ = *Rank* data *training* ke-*a* pada variabel ke-*q*
$R_q(b)$ = *Rank* data *testing* ke-*b* pada variabel ke-*q*
maks(R_q) = *Rank* maksimum variabel ke-*q*
min(R_q) = *Rank* minimum variabel ke-*q*
min(x_q) = Nilai minimum dari variabel *q*
maks(x_q) = Nilai maksimum dari variabel *q*
d(a,b) = Jarak Gower antara data *training* ke-*a* dengan data *testing* ke-*b*

2.4 k-Fold Cross Validation

Cross Validation adalah teknik validasi dengan membagi data secara acak ke dalam k bagian di mana masing-masing bagian akan dilakukan proses klasifikasi. *k-fold cross validation* merupakan bagian dari pengujian *cross validation* yang berfungsi untuk menilai kinerja proses dengan membagi sampel data secara acak dan dikelompokkan sebanyak nilai *k-fold* (Han dkk, 2012).

Secara keseluruhan, 5 atau 10-*fold cross validation* sama-sama direkomendasikan dan disepakati bersama. Jumlah data di dalam satu *subset* data dapat dihitung menggunakan persamaan sebagai berikut,

$$B = \frac{tr}{k} \tag{7}$$

Keterangan :
B = Banyak data dalam setiap *subset* data
tr = Banyaknya data *training*
k = Nilai *k-fold cross validation*

2.5 Pengukuran Tingkat Akurasi

Menurut Satria dkk (2019), pengukuran tingkat akurasi klasifikasi dapat dilakukan dengan menghitung kebenaran klasifikasi. Perhitungan tingkat akurasi dapat ditunjukkan pada persamaan berikut,

$$\text{Akurasi} = \frac{\text{Banyak data uji benar klasifikasi}}{\text{Banyak data}} \times 100\% \tag{8}$$

Perhitungan akurasi hasil klasifikasi untuk *subset* data pada setiap nilai K menggunakan persamaan berikut,

$$A_{(g,K)} = \frac{l}{m} \times 100\% \tag{9}$$

Perhitungan rata-rata akurasi hasil klasifikasi untuk setiap nilai K menggunakan persamaan berikut,

$$\text{Akurasi(K)} = \frac{\text{Total akurasi pada nilai K}}{\text{Banyaknya } subset} \times 100\% \tag{10}$$

Keterangan :
$A_{(g,K)}$ = Akurasi untuk *subset* data ke-*g* pada himpunan K ketetanggaan terdekat
l = Banyak data klasifikasi benar dalam satu *subset* data *testing*
m = Banyak data dalam satu *subset* data *testing*
K = Banyak tetangga terdekat
Akurasi(K) = Rata-rata akurasi pada nilai K

2.6 Penyakit Hipertensi

Penyakit Hipertensi merupakan penyakit kardiovaskular yang berarti peningkatan abnormal pada tekanan darah baik sistolik maupun diastolik. Seseorang dapat dikatakan menderita hipertensi jika tekanan darah sistolik/diastolik lebih dari 140/90 mmHg (tekanan darah normal 120/80 mmHg). Hipertensi sangat terkait dengan perubahan gaya hidup, konsumsi makanan yang berlemak tinggi, kolesterol, kurangnya aktivitas olahraga dan stres (Herwati & Wiwi, 2011).

Penderita penyakit ini tidak akan merasakan tanda-tanda seperti penyakit pada umumnya dikarenakan sulit untuk dideteksi. Namun ketika dilakukan pemeriksaan terkait dengan penyakit yang berhubungan dengan hipertensi seperti stroke dan diabetes baru akan terdeteksi.

Faktor risiko hipertensi merupakan kebiasaan individu yang lebih umum dialami oleh penderita daripada orang lain yang normal. Atribut individu tersebut dapat berupa umur, jenis kelamin, atau riwayat penyakit tertentu. Sedangkan kebiasaan yang dapat menjadi faktor risiko dapat berupa kebiasaan merokok, penyalahgunaan narkoba, asupan makanan, dan kebiasaan olahraga (Yogiantoro, 2009).

3. Bahan dan Metode

Populasi penelitian ini adalah seluruh pasien rawat jalan UPTD Puskesmas Sempaja Kota Samarinda. Sampel pada penelitian ini adalah pasien rawat jalan UPTD Puskesmas Sempaja Kota Samarinda bulan Februari - Mei 2022.

Variabel penelitian ini yaitu status hipertensi sebagai variabel terikat dan usia, jenis kelamin, diabetes mellitus, serta penyakit jantung sebagai variabel bebas. Variabel penelitian dapat dilihat pada Tabel

p-ISSN 2085-7829
e-ISSN 2798-3455

1.

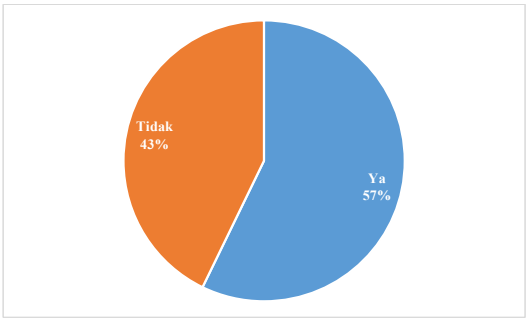
Tabel 1. Variabel Penelitian

Variabel Penelitian	Definisi	Tipe Data
Status Hipertensi	Status hipertensi pasien dengan kategori Ya dinotasikan sebagai 1 dan tidak dinotasikan sebagai 2	Nominal
Usia (Tahun)	Usia pasien	Rasio
Jenis Kelamin	Jenis kelamin pasien dengan kategori Laki-Laki dinotasikan sebagai L dan Perempuan dinotasikan sebagai P.	Nominal
Diabetes Melitus	Riwayat penyakit diabetes melitus pasien dengan kategori Ya dan Tidak.	Nominal
Penyakit Jantung	Penyakit jantung pasien dengan kategori Ya dan Tidak.	Nominal

4. Hasil dan Pembahasan

4.1 Deskripsi Data Penelitian

Data yang digunakan dalam penelitian ini yaitu 278 data pasien rawat jalan di UPTD Puskesmas Sempaja Kota Samarinda yang merupakan salah satu fasilitas pelayanan kesehatan masyarakat di Kota Samarinda. Data 278 pasien rawat jalan pada bulan Februari - Mei 2022 dengan status hipertensi, usia, jenis kelamin, diabetes mellitus, dan penyakit jantung. Deskripsi data status hipertensi dapat dilihat pada Gambar 1.



Gambar 1. Status Hipertensi

Gambar 1 menunjukkan bahwa status hipertensi 278 pasien rawat jalan UPTD Puskesmas Sempaja Kota Samarinda terdapat 119 pasien yang tidak hipertensi dengan persentase 42,8% dan 159 pasien berstatus hipertensi dengan persentase 57,2%.

4.2 Pembagian Data *Training* dan Data *Testing*

Data diacak dan dibagi menjadi data *training* dan data *testing* menggunakan tiga proporsi, yaitu 50% : 50%, 70% : 30%, dan 90% : 10%. Data yang berada pada urutan pertama akan digunakan menjadi data *training* dan sisanya akan menjadi data *testing*. Proporsi 70% : 30% digunakan sebagai contoh dalam menguraikan perhitungan analisis. Penggunaan proporsi 70% : 30% menandakan bahwa 70% dari hasil pengacakan data pada urutan pertama akan menjadi data *training* dan 30% sisanya akan menjadi data *testing*. Perhitungan pembagian banyaknya data *training* dengan menggunakan persamaan (1) sebagai berikut,

$$tr = \frac{70}{100} \times 278$$
$$= 194{,}6 \approx 195$$

Adapun perhitungan pembagian banyaknya data *testing* dengan menggunakan persamaan (2) sebagai berikut,

$$te = 278 - 195$$
$$= 83$$

Berdasarkan hasil perhitungan banyaknya data *training* yang diperoleh, bahwa 195 data pertama hasil pengacakan data untuk masing-masing variabel penelitian berfungsi sebagai data *training* yang akan digunakan pada metode K-NN.

4.3 Penentuan Nilai *k* pada *k-Fold Cross Validation*

p-ISSN 2085-7829
e-ISSN 2798-3455

K-fold cross validation merupakan metode untuk mengetahui rata-rata keberhasilan dengan cara melakukan perulangan tiap *subset* data dari pengacakan data yang diperoleh. Penggunaan *k* dilakukan dalam pembagian himpunan secara acak yang akan menjadi *subset* data.

Dalam penelitian ini digunakan 10-*fold cross validation*, maka jumlah *fold* yang diperlukan yaitu sebanyak 10-*fold*. Penggunaan 10-*fold* tersebut akan membentuk *subset* data sedemikian sehingga jumlah *subset* yang digunakan sebanyak 10 *subset* data. Sebagai contoh, pada *fold* ke-1 terdapat 9 kombinasi *subset* data yang berbeda sebagai data *training* dan sisanya terdapat 1 *subset* data sebagai data *testing*, selanjutnya proses *training* dan *testing* dilakukan sampai *fold* ke-10.

4.4 Penentuan Banyaknya Data dalam *Subset* Data *Training*

Perhitungan jumlah data dalam *subset* hanya menggunakan banyaknya data *training* dari proporsi yang ditentukan yaitu proporsi 70% yakni dengan jumlah data *training* sebanyak 195 data serta nilai *k-fold cross validation* yang ditentukan yaitu *10-fold cross validation*. Berikut perhitungan jumlah data dalam *subset* yaitu menggunakan Persamaan (7).

$$B = \frac{195}{10}$$
$$= 19{,}5 \approx 19 \text{ atau } 20$$

Berdasarkan perhitungan jumlah data untuk setiap *subset* yang nantinya digunakan pada analisis metode K-NN yaitu sebanyak 19 atau 20 data pengamatan.

4.5 Perhitungan Jarak Ketetanggaan antara Data *Testing* terhadap Data *Training* Pada *Subset* Data

Pada penelitian ini menggunakan *k-fold cross validation* yaitu *10-fold cross validation* sedemikian sehingga 1 *subset* akan berlaku sebagai data *testing* dan *subset* lainnya sebagai data *training*. Setiap *subset* memiliki giliran untuk dijadikan sebagai data *testing*. Sebagai contoh, ketika *subset* ke-1 sebagai data *testing* maka *subset* ke-2 hingga *subset* ke-10 sebagai data *training* dan seterusnya hingga *subset* ke-10 sebagai data *testing* dan *subset* ke-1 hingga *subset* ke-9 sebagai data *training*.

Pada penelitian ini menggunakan perhitungan jarak antara data *testing* ke-1 yakni data ke-1 pada *subset* 1 (Pasien ke-83) dengan data *training* ke-1 yakni data ke-1 pada *subset* 2 (Pasien ke-124) sebagai contoh dalam melakukan proses perhitungan jarak ketetanggaan menggunakan jarak Gower. Perhitungan jarak Gower dilakukan dengan menghitung jarak data *testing* terhadap data *training* pada setiap variabel *subset* data.

Pada variabel X_1 dengan skala rasio, maka perhitungan dilakukan menggunakan persamaan (5) sebagai berikut,

$$d_1(1,1) = \frac{|x_{11} - y_{11}|}{\text{maks}(x_1) - \min(x_1)}$$
$$= \frac{|22 - 39|}{\text{maks}(44) - \min(16)}$$
$$= \frac{17}{28}$$
$$= 0{,}6071$$

Pada variabel X_2 dengan skala nominal, maka perhitungan dilakukan menggunakan persamaan (3) sebagai berikut:

$$d_4(1,1) = x_{14} = x_{14}$$
$$= 0$$

Perhitungan jarak Gower dilakukan hingga *subset* ke-10 sebagai data *testing* dan *subset* ke-1 hingga *subset* ke-9 sebagai data *training* yaitu data *testing* ke-19 yakni data ke-19 pada *subset* 10 (Pasien ke-101) dengan data *training* ke-176 yakni data ke-19 pada *subset* 9 (Pasien ke-17).

4.6 Pengklasifikasian Berdasarkan Nilai K pada Setiap *Subset* Data

Setelah dilakukan perhitungan jarak ketetanggaan menggunakan jarak Gower antara data *training* dengan data *testing* pada *subset* data. Selanjutnya, dilakukan pengurutan jarak Gower dari jarak yang paling dekat sampai dengan jarak yang paling jauh pada setiap *subset* data. Setelah diperoleh hasil pengurutan jarak Gower, ditetapkan nilai K tetangga terdekat menggunakan K yang bernilai ganjil yaitu 3, 5, 7, dan 9 dikarenakan jumlah kategori pada variabel terikat dalam penelitian ini berjumlah genap yakni kategori Ya dan Tidak. Pemilihan kategori pada variabel terikat dilakukan dengan tahap *voting*. Hal ini bertujuan agar pada saat *voting* tetangga terdekat tidak menimbulkan hasil *voting* yang seri atau sama yang mengakibatkan peneliti tidak mendapatkan jawaban hasil klasifikasinya.

Pada penelitian ini digunakan data *testing* ke-1 yaitu data ke-1 pada *subset* ke-1 (Pasien ke-83) sebagai contoh dalam melakukan penentuan hasil klasifikasi data *testing*. Jumlah data *testing* pada *subset* 1 yakni terdapat 20 data *testing* dengan *subset* 2 hingga *subset* 10 sebagai data *training* berjumlah 175 data *training*. Berikut adalah data *rank* jarak Gower untuk data *testing* ke-1 yakni data ke-1 pada *subset* 1 (Pasien ke-83) yang dapat dilihat pada Tabel 2.

p-ISSN 2085-7829
e-ISSN 2798-3455

Tabel 2. Hasil Pengurutan Jarak Gower 175 Data *Training* Terhadap Data *Testing* ke-1 dari *Subset* 1 (Pasien ke-83)

Data *Training*		Batas K-NN	Hasil Klasifikasi Data *Testing* ke-1
Pasien	**Klasifikasi**		
150	Tidak		
217	Ya		
78	Ya	3	Ya
212	Ya		
178	Tidak	5	Ya
34	Tidak		
82	Ya	7	Ya
⋮	⋮	⋮	⋮
103	Ya		

Selanjutnya dilakukan hal yang sama data *testing* ke-2 yakni data ke-2 pada *subset* 1 (Pasien ke-177), data *testing* ke-3 yakni data ke-3 pada *subset* 1 (Pasien ke-176), sampai dengan data *testing* ke-19 yakni data ke-19 pada *subset* 10 (Pasien ke-101) sehingga mendapatkan hasil klasifikasi data *testing* tiap *subset* data.

4.7 Pemilihan K Optimum dan Proporsi Terbaik Berdasarkan *k-Fold Cross Validation*

Pemilihan nilai K optimum dan proporsi terbaik dapat dilakukan dengan melihat nilai akurasi yang dihasilkan. Sebelum melakukan pemilihan, hasil klasifikasi kategori status hipertensi untuk semua data *testing* dari 10 *subset* data dengan menggunakan nilai K ganjil akan dibandingkan dengan klasifikasi pada data aslinya. Kemudian dilakukan perhitungan akurasi dari masing-masing *subset* dengan menggunakan Persamaan (9). Pada penelitian ini ketika *subset* 1 menjadi data *testing* sebagai contoh dalam melakukan perhitungan akurasi. Perbandingan akurasi hasil prediksi klasifikasi dapat dilihat pada Tabel 3.

Tabel 3. Perbandingan Akurasi Hasil Klasifikasi pada *Subset* 1

No. Sampel	**Klasifikasi berdasarkan Nilai K pada *Subset* 1**				**Klasifikasi Pada Data Asli (*Y*)**
	3	**5**	**7**	**9**	
83	Ya	Ya	Ya	Ya	Ya
177	Ya	Ya	Tidak*	Tidak*	Ya
176	Ya*	Ya*	Ya*	Ya*	Tidak
⋮	⋮	⋮	⋮	⋮	⋮
264	Ya	Ya	Ya	Ya	Ya
Prediksi Benar	9	12	10	10	
$A_{(1,K)}$	0,45	0,6	0,5	0,5	

Berdasarkan Tabel 3 dapat dilihat bahwa kategori dengan tanda (*) memiliki perbedaan klasifikasi dengan klasifikasi pada data aslinya. Hasil klasifikasi yang benar pada masing-masing nilai K akan dihitung jumlahnya kemudian menghitung akurasi hasil klasifikasi dari *subset* 1. Semakin banyak jumlah prediksi klasifikasi yang tepat maka semakin baik juga nilai K yang digunakan untuk memperoleh nilai K optimum dalam mengklasifikasikan kategori status hipertensi. Langkah-langkah tersebut akan dilakukan juga pada *subset* 2, *subset* 3, hingga *subset* 10 untuk mendapatkan nilai akurasi pada setiap *subset* data.

Setelah diperoleh akurasi pada setiap *subset*, kemudian dilakukan perhitungan rata-rata keseluruhan nilai akurasi berdasarkan nilai K dengan menggunakan Persamaan (10). Pada penelitian ini digunakan batas 3 tetangga terdekat (3-NN) sebagai contoh dalam perhitungan nilai akurasi klasifikasi 3-NN yaitu sebagai berikut,

$$\text{Akurasi}(3) = \frac{0,45 + 0,6 + 0,4 + \ldots + 0,53}{10} \times 100\%$$
$$= 56,40\%$$

Adapun hasil perhitungan akurasi prediksi klasifikasi untuk semua *subset* dengan menggunakan 10-

p-ISSN 2085-7829
e-ISSN 2798-3455

fold cross validation dapat dilihat pada Tabel 4 sebagai berikut,

Tabel 4. Akurasi Prediksi Klasifikasi Proporsi 70% dengan *10-Fold Cross Validation*

Subset	**Nilai K**			
	3	**5**	**7**	**9**
1	0,45	0,6	0,5	0,5
2	0,6	0,6	0,6	0,55
3	0,4	0,55	0,65	0,5
4	0,55	0,65	0,55	0,5
5	0,7	0,7	0,8	0,8
6	0,47	0,47	0,42	0,53
7	0,58	0,63	0,63	0,68
8	0,68	0,74	0,68	0,68
9	0,68	0,74	0,63	0,58
10	0,53	0,58	0,63	0,53
Rata - Rata Akurasi (%)	56,40	62,60	60,90	58,50

Berdasarkan Tabel 4 dapat dilihat bahwa nilai akurasi pada 3 tetangga terdekat (3-NN) yaitu sebesar 56,40%, pada 5 tetangga terdekat (5-NN) yaitu sebesar 62,60% dan seterusnya hingga 9 tetangga terdekat (9-NN) yaitu sebesar 58,50%.

Semua tahapan klasifikasi diterapkan juga pada 2 proporsi lainnya yakni pada proporsi 50% : 50% dan 90% : 10%. Setelah didapat akurasi masing-masing proporsi kemudian, akurasi tersebut dibandingkan dan dipilih nilai akurasi tertinggi. Semakin besar jumlah klasifikasi yang benar maka semakin besar kemungkinan nilai K tersebut merupakan K optimum. Berikut ini merupakan nilai akurasi dengan 10-*fold cross validation* dari beberapa proporsi data *training* dan data *testing* yang dapat dilihat pada Tabel 5.

Tabel 5. Akurasi Prediksi Klasifikasi Berdasarkan Nilai K

Perbandingan data *Training* dan *Testing*	Nilai akurasi dengan 10-*fold cross validation*			
	K = 3	K = 5	K = 7	K = 9
50% : 50%	54%	49,8%	51,1 %	50,4%
70% : 30%	56,4 %	**62,6%**	60,9 %	58,5%
90% : 10%	55,2 %	58,4 %	59,2 %	60,8%

Berdasarkan Tabel 5 dapat dilihat nilai akurasi yang dicetak tebal merupakan nilai akurasi tertinggi. Nilai akurasi tertinggi terdapat pada proporsi 70% : 30% yakni pada batas 5 tetangga terdekat (5-NN) dengan nilai akurasinya sebesar 62,6%. Sehingga, nilai K optimum yang digunakan pada tahap analisis selanjutnya yaitu K = 5 yang berasal dari proporsi 70% : 30% dengan 195 data *training* dan 83 data *testing*.

Jika diperhatikan nilai akurasi tertinggi dalam penelitian ini masih tergolong rendah karena hanya 62,6%. Hal ini bisa disebabkan karena distribusi data tidak sama rata yaitu jumlah kelas data pasien yang tidak hipertensi lebih sedikit dibanding dengan jumlah kelas data pasien yang hipertensi. Peneliti selanjutnya dapat menggunakan pengembangan dari algoritma K-NN yaitu *Neighbor Weighted K-Nearest Neighbor* karena cocok untuk diimplementasikan kepada data yang tidak berdistribusi secara rata.

5. Kesimpulan

Berdasarkan hasil analisis data dan pembahasan diperoleh kesimpulan bahwa persentase akurasi pengklasifikasian dengan metode K-NN dalam pengklasifikasian status hipertensi pasien di UPTD Puskesmas Sempaja Kota Samarinda yaitu 62,60% dengan nilai K optimal dalam pengklasifikasian status hipertensi pasien menggunakan 10-*fold cross validation* yaitu K = 5 atau sebanyak 5 tetangga terdekat. Proporsi terbaik dalam klasifikasi status hipertensi pasien UPTD Puskesmas Sempaja Kota Samarinda adalah 70%:30%.

p-ISSN 2085-7829
e-ISSN 2798-3455

Referensi

Annur, H. (2018). Klasifikasi Masyarakat Miskin Menggunakan Naive Bayes. *ILKOM Jurnal Ilmiah*, 10(2), 160-165.

Bellazi, R. & Zupan, B. (2008). Predictive Data Mining in Clinical Medicine: Current Issues And Guidlines. *International Journal of Medical Informatics*, 7, 81-97.

Han, J. W., Kamber, M. & Pei. (2012). *Data Mining Concepts and Techniques*. San Fransisco: Morgan Kaufmann Publishers.

Herwati & Wiwi, S., (2011). Terkontrolnya Tekanan Darah Penderita Hipertensi Berdasarkan Pola Diet Kebiasaan Olah Raga di Padang Tahun 2011. *Jurnal Kesehatan Masyarakat*, 8-14.

Krisandi, N., Helmi, & Prihandoso, B. (2013). Algoritma K-Nearest Neighbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada PT. MINAMAS Kecamatan Parindu. *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, 02, 33-38.

Nyoman, W., (2016). *Klasifikasi Penyakit Hipertensi Menggunakan Algoritma Naïve Bayes*. Yogyakarta: Universitas Sanata Dharma.

Prasath, V. B., Alfeith, H.A., Lasassmeh, O., & Hassanat, A.B. (2017). *Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier - A Review*, 1-50.

Prasetyo, E. (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Penerbit Andi.

Satria, A., Marji, & Ratnawati, D. E. (2019). Klasifikasi Jenis Kanker Berdasarkan Struktur Protein Menggunakan Metode Neighbor Weighted K-Nearest Neighbor (NWKNN). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(4), 3617-3624.

Sidabutar, R. P & Wiguno, P. (2009). *Hipertensi Esensial Penyakit Dalam*. Jakarta: FK-UI.

Sriwahyuni, E. & Khumaini, A. H. (2018). Sistem Kewaspadaan Dini dan Respon Hipertensi. *Berita Kedokteran Masyarakat*, 34(11), 16-18.

Tjiptoherijanto, P. (1994). *Ekonomi Kesehatan*. Jakarta: Rineka Cipta.

Yogiantoro, M. (2009). *Hipertensi Esensial Ilmu Penyakit Dalam*. Jakarta: Interna.

Yustanti, W. (2012). Algoritma K-Nearest Neighbor Untuk Memprediksi Harga Jual Tanah. *Jurnal Matematika,Statistika & Komputasi*, 9(1), 57-68.