Penerapan Metode Klasifikasi *K-Nearest Neighbor* (Studi Kasus : Data Status Gizi Balita di Puskesmas Baqa Samarinda Seberang)

Muzizah Annabaa' Aulia[1*], Rito Goejantoro[2], Memi Nor Hayati[3]

[1,2]Laboratorium Statistika Komputasi, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman, Indonesia

[3]Laboratorium Statistika Terapan, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman, Indonesia

Corresponding author: muzizah09@gmail.com

Abstrak. Klasifikasi adalah suatu proses mengamati objek data yang bertujuan untuk membedakan kelompok-kelompok yang tersedia ke dalam kelompok tertentu. Metode *K-Nearest Neighbor* (K-NN) merupakan metode *supervised* yang digunakan untuk klasifikasi (dengan peubah *output* atau peubah dependen berupa kategori). Prinsip kerja metode ini ialah menemukan jarak terdekat antara data yang akan dievaluasi dengan nilai *K* tetangga (*neighbor*) terdekatnya dalam data *testing*, dari nilai *K* tetangga terdekat yang terpilih kemudian akan dilakukan pemungutan kelas terbanyak atau *voting* kelas dari *K* tetangga terdekat. Status gizi balita merupakan keadaan gizi anak balita umur 0-59 bulan yang ditentukan dengan metode antropometri, berdasarkan indeks berat badan menurut umur (BB/U), tinggi badan menurut umur (TB/U), dan berat badan menurut tinggi badan (BB/TB). Adapun tujuan dari penelitian ini adalah untuk memperoleh hasil klasifikasi dari status gizi balita di Puskesmas Baqa Samarinda Seberang menggunakan metode klasifikasi K-NN. Berdasarkan hasil pengukuran untuk klasifikasi pada status gizi balita di Puskesmas Baqa Samarinda Seberang dengan perhitungan akurasi serta matriks konfusi didapatkan akurasi tertinggi pada metode K-NN sebesar 90,57% pada nilai K= 3,K= 5, K= 7 dan K= 9 dengan proporsi data *training* 90% dan data *testing* 10%.

Kata Kunci: klasifikasi, *k-nearest neighbor*, status gizi balita

1 PENDAHULUAN

Adanya peningkatan pemakaian internet di seluruh dunia yang berkembang sangat pesat berakibat pula pada terjadinya peningkatan data dalam jumlah yang sangat besar. Agar dapat memberikan manfaat dan dapat digunakan untuk dijadikan sebagai patokan dalam menentukan sebuah keputusan, maka sebaiknya data harus diproses dan dikumpulkan dahulu. Dalam proses pengumpulan data biasanya dikenal dengan istilah *data mining*. Menurut [1], *data mining* merupakan proses pencarian pola dan relasi-relasi yang tersembunyi dalam data yang besar untuk melakukan estimasi, prediksi, *association rule*, *clustering*, deskripsi, visualisasi dan klasifikasi.

Klasifikasi merupakan salah satu metode yang paling sering digunakan di *data mining* [2]. Dalam melakukan suatu klasifikasi, biasanya digunakan algoritma klasifikasi, adapun algoritma yang sering digunakan dalam proses pengklasifikasian ialah *K-Neirest Neighbor*. K-NN merupakan sebuah metode yang dilakukan untuk mengklasifikasikan objek berdasarkan dengan data pembelajaran (*neighbor*) yang jaraknya berdekatan dengan objek [3]. Pada penelitian ini penulis menggunakan algoritma klasifikasi yaitu K-NN. Adapun pertimbangan dari dipilihnya metode K-NN sebagai metode klasifikasi yang digunakan dalam pengklasifikasian dikarenakan tingkat akurasi dari metode klasifikasi ini relatif tinggi. Hal ini dapat dibuktikan pada penelitian sebelumnya yang dilakukan oleh Rahmaulidyah dkk (2021) [4] mengenai perbandingan metode klasifikasi *naïve* Bayes dan *K-Neirest Neighbor* pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Samarinda Ulu yang menunjukkan hasil kesalahan klasifikasi dalam memprediksi sebesar 19,51%. Penelitian yang dilakukan oleh Mustaghfiroh dkk (2022) [5] mengenai klasifikasi pasien Covid-19 di Indonesia menggunakan metode K-NN dengan nilai akurasi mencapai 97,76%. Sedangkan, penelitian yang dilakukan oleh Saeroni dkk (2020) [6] mengenai klasifikasi tingkat kelancaran nasabah dalam membayar premi dengan menggunakan metode *K-Nearest Neighbor* dan Analisis Diskriminan Fisher (Studi kasus: Data Nasabah PT. Prudential Life Samarinda Tahun 2019) menunjukkan bahwa hasil kesalahan klasifikasi dalam memprediksi kelas menggunakan APER sebesar 15%.

Metode klasifikasi K-NN dapat dimanfaatkan dan diaplikasikan ke dalam berbagai bidang, salah satunya adalah dalam bidang kesehatan masyarakat. Dalam ruang lingkupnya yang luas kesehatan masyarakat terbagi ke dalam beberapa lingkup salah satunya ialah gizi. Menurut [7], gizi merupakan suatu proses organisme menggunakan makanan yang dikonsumsi secara normal melalui proses digesti, absorpsi, transportasi, penyimpanan, metabolisme dan pengeluaran zat-zat tidak digunakan untuk mempertahankan kehidupan. Sedangkan status gizi merupakan ekspresi dan keadaan keseimbangan dalam bentuk variabel tertentu atau perwujudan dari *nutriture* dalam bentuk variabel tertentu.

Masalah gizi yang biasanya terjadi antara lain yaitu gizi kurang dan gizi buruk. Usia di bawah lima tahun merupakan usia di mana tahapan perkembangan terjadi dan rentan akan penyakit yang disebabkan oleh kekurangan maupun kelebihan nutrisi [8]. Dalam mencegah dan mengurangi kenaikan pada permasalahan gizi, pemerintah melakukan berbagai tindakan untuk mempercepat penurunannya. Salah satu langkah yang dilakukan adalah dengan melakukan pemantauan pertumbuhan balita pada setiap Puskesmas maupun Posyandu. Selain

menjadi layanan kesehatan bagi masyarakat, Puskesmas juga berfungsi dalam membina masyarakat dalam meningkatkan kemampuan hidup sehat, pengumpulan serta pendataan, melaksanakan komunikasi serta informasi kepada masyarakat dalam bidang kesehatan. Dalam hal pengumpulan dan pendataan diperlukan adanya suatu sistem klasifikasi untuk mempermudah penyajian informasi mengenai gizi balita yang dibutuhkan. Berdasarkan latar belakang tersebut, maka penulis tertarik untuk melakukan penelitian dengan judul "Penerapan Metode Klasifikasi K-NN pada Data Status Gizi Balita di Puskesmas Baqa Samarinda Seberang".

2 TINJAUAN PUSTAKA

2.1 *Data Mining*

Data mining adalah suatu istilah untuk menemukan segala sesuatu yang tersembunyi di dalam *database*. *Data mining* adalah proses menggunakan teknik statistik, matematika, kecerdasan buatan, *machine learning* untuk mengutip dan mengidentifikasi informasi yang bermanfaat dan hal-hal yang berkaitan dengan ilmu basis data besar [9].

Data mining merupakan proses dalam menemukan hal-hal yang berhubungan dengan pola dan kecenderungan dengan mengamati sekumpulan besar data yang tersimpan, dengan bantuan teknik pengenalan pola [10].

2.2 Proses *Data Mining*

Secara sistematis, terdapat tiga langkah utama dalam *data mining* di antaranya sebagai berikut [11]:

1. Eksplorasi/pemrosesan awal data
   Eksplorasi atau pemrosesan awal data terdiri dari 'pembersihan' data, normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan *subset* fitur dan sebagainya.
2. Membangun model dan melakukan validasi terhadapnya
   Membangun model dan melakukan validasi terhadapnya berarti melakukan analisis berbagai model dan memilih model dengan kinerja prediksi terbaik. Dalam langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis klaster, deteksi anomali, analisis asosiasi, analisis pola sekuensial, dan sebagainya.
3. Penerapan
   Penerapan merupakan menerapkan model pada data yang baru untuk menghasilkan perkiraan atau prediksi masalah yang diinvestigasi.

2.3 Kelompok *Data Mining*

Metode *data mining* terbagi dalam enam kelompok berdasarkan tugas dengan fungsi yang dilakukan. Kelompok-kelompok tersebut diuraikan sebagai berikut [10]:

1. Deskripsi
   Dalam metode ini digunakan peneliti untuk menganalisis pola dan kecenderungan yang belum terlihat di dalam data.
2. Klasifikasi
   Pada metode klasifikasi digunakan dalam proses mengkategorikan variabel atau kelas data.

3. Estimasi
   Sama halnya dengan klasifikasi, estimasi merupakan proses mengkategorikan atau membedakan variabel target, namun lebih mengarah ke numerik dibandingkan kategorik.
4. Prediksi
   Prediksi hampir sama dengan klasifikasi dan estimasi, tetapi nilai dari hasil prediksi akan ada di masa mendatang.
5. Klaster
   Klaster merupakan pengelompokan sejumlah data yang memiliki kemiripan dalam kelompok-kelompok data. Letak perbedaan antara klaster dengan klasifikasi yaitu tidak terdapat variabel target dalam pengklasteran.
6. Asosiasi
   Teknik yang digunakan untuk mencari hubungan antara karakteristik tertentu dalam satu waktu.

2.4 Klasifikasi

Klasifikasi adalah suatu proses mengamati objek data yang bertujuan untuk membedakan kelompok-kelompok yang tersedia ke dalam kelompok tertentu. Klasifikasi memiliki dua fungsi pokok, yaitu pembangunan model prototipe untuk disimpan sebagai memori dan melakukan pengenalan atau prediksi suatu objek data lain untuk mengetahui penempatan kelompok dari objek data tersebut dalam model yang telah disimpannya [11].

Klasifikasi adalah sebuah proses untuk mencari model atau fungsi yang menjelaskan serta membedakan konsep atau kelas dari data, dengan tujuan untuk menggunakan model dan melakukan prediksi dari kelas suatu objek di mana tidak diketahui label dari kelas tersebut [1].

2.5 Metode *K-Nearest Neighbor*

Metode K-NN sendiri pertama kali diperkenalkan oleh Evelyn Fix dan Joseph Hodges pada tahun 1951 dan kemudian dikembangkan oleh Thomas Cover. Metode K-NN merupakan metode *supervised* yang digunakan untuk klasifikasi (dengan peubah *output* atau peubah dependen berupa kategori). *Supervised* (pembelajaran yang diawasi) merupakan proses pembelajaran dari data *training* yang memandu dan mengoreksi pembelajaran sampai algoritma mencapai tingkat kinerja yang dapat diterima [1].

Prinsip kerja metode ini ialah menemukan jarak terdekat antara data yang akan dievaluasi dengan nilai *K* tetangga (*neighbor*) terdekatnya dalam data *testing*, dari nilai *K* tetangga terdekat yang terpilih kemudian akan dilakukan pemungutan kelas terbanyak atau *voting* kelas dari *K* tetangga terdekat. Jika terdapat kelas yang memiliki jumlah suara tetangga terbanyak, maka selanjutnya akan diberikan label kelas hasil prediksi pada data *testing* tersebut [12]. Jauh atau dekatnya jarak titik dengan tetangganya bisa dihitung dengan menggunakan jarak Euclid yang dipresentasikan sebagai berikut [11]:

$$d(x_{ik}, x^*_{jk}) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x^*_{jk})^2} \tag{2.1}$$

di mana :
$d(x_{ik}, x^*_{jk})$: jarak Euclid data *training* ke-i dengan data *testing* ke-j
x_{ik} : nilai variabel bebas ke-k dari data *training* ke-i, $i = 1,2,\dots,n_{tr}$
x^*_{jk} : nilai variabel bebas ke-k dari data *testing* ke-i, $j = 1,2,\dots,n_{tes}$
p : banyaknya variabel bebas

2.6 Standardisasi Data

Jika jarak Euclid semakin kecil, maka semakin mirip kasus atau objek tersebut. Akan tetapi, jarak Euclid sangat sensitif terhadap ukuran sampel dan besarnya varian. Jika objek yang diteliti memiliki varian yang sangat berbeda, maka jarak Euclid menjadi tidak akurat. Oleh sebab itu, perlu dilakukan standardisasi terhadap variabel penelitian [13].

$$Z_{rt} = \frac{X_{rt} - \bar{X}_t}{S_t} \tag{2.2}$$

dengan:
Z_{rt} : data hasil standardisasi observasi ke-r variabel ke-t
X_{rt} : observasi ke-r variabel ke-t
$\bar{X}_t$: rata-rata variabel ke-t
S_t : simpangan baku variabel ke-t

2.7 Data *training* dan data *testing*

Data *training* digunakan oleh algoritma klasifikasi untuk membentuk sebuah model *classifier*. Model ini merupakan representasi pengetahuan yang akan digunakan untuk memprediksi kelompok dari data baru yang belum pernah ada. Data *testing* digunakan untuk mengukur sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar. Data yang ada pada data *testing* seharusnya tidak boleh ada pada data *training* sehingga dapat diketahui apakah model *classifier* sudah tepat atau belum dalam melakukan klasifikasi [2]. Jumlah data *training* dan data *testing* dapat dihitung menggunakan Persamaan (2.3) dan Persamaan (2.4) sebagai berikut.

$$\text{Jumlah data } \textit{training} = \text{proporsi data } \textit{training} \times n \tag{2.3}$$

$$\text{Jumlah data } \textit{testing} = n - \text{jumlah data } \textit{training} \tag{2.4}$$

dengan :
n : jumlah seluruh data (data *training* + data *testing*)

2.8 Evaluasi Ketepatan Hasil Klasifikasi

Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua *set* data dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja suatu sistem tidak bisa 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (*confusion matrix*) [14].

Matriks konfusi merupakan tabel pencatatan hasil kerja klasifikasi. Tabel 2.1 merupakan contoh matriks konfusi yang melakukan klasifikasi masalah biner (dua kelas), hanya ada dua kelas yaitu kelas 0 dan kelas 1. Setiap sel f_{ab} dalam matriks menyatakan jumlah *record* atau data dari kelas a yang hasil prediksinya

masuk kelas b. Misalnya sel f_{11} adalah jumlah data dalam kelas 1 yang secara benar dipetakan ke kelas 1, dan f_{10} adalah data dalam kelas 1 yang dipetakan secara salah ke kelas 0.

Tabel 2.1 Matriks Konfusi untuk Klasifikasi Dua Kelas

f_{ab}		Kelas hasil prediksi (b)	
		Kelas = 1	Kelas = 0
Kelas asli (a)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Berdasarkan isi matriks konfusi, kita dapat mengetahui jumlah data dari masing-masing kelas yang diprediksi secara benar, yaitu $(f_{11} + f_{00})$ dan data yang diklasfikasi secara salah, yaitu $(f_{10} + f_{01})$. Kuantitas matriks konfusi dapat diringkas menjadi dua nilai, yaitu akurasi dan laju error. Dengan megetahui jumlah data yang yang diklasifikasi secara benar, kita dapat mengetahui akurasi hasil prediksi, dan mengetahui jumlah data yang diklasifikasi secara salah, kita dapat mengetahui laju error dari prediksi yang dilakukan. Untuk menghitung akurasi dapat digunakan rumus sebagai berikut [14]:

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Jumlah prediksi yang dilakukan}} \times 100\%$$

$$= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \times 100\% \quad (2.5)$$

Uji kestabilan dilakukan untuk menguji apakah pengalokasian dari tiap sampel dalam kelompok relatif stabil atau tidak sebagai akibat adanya perubahan perbedaan jumlah sampel yang diteliti. Adapun uji yang dapat dilakukan adalah dengan menghitung nilai *Press's Q* yang diformulasikan sebagai berikut:

$$Press's\ Q = \frac{[n - (mq)]^2}{n(q-1)} \quad (2.6)$$

di mana:
n : jumlah sampel secara keseluruhan
m : jumlah pengamatan yang tepat diklasifikasikan
q : banyaknya kelompok

Pengklasifikasian dikatakan akurat apabila nilai *Press'S Q* lebih besar daripada nilai kritis yang diambil dari tabel *chi square* $(\mathcal{X}^2_{\alpha,1})$ dengan derajat bebas bernilai satu dan tingkat keyakinan sesuai dengan yang diinginkan. [15].

2.9 Status Gizi Balita

Status gizi merupakan hasil akhir dari keseimbangan antara nutrisi atau asupan yang dikonsumsi dan diserap atau dimasukkan ke dalam tubuh (*nutrient input*) dengan kebutuhan tubuh (*nutrient output*) akan zat gizi tersebut [16].

Untuk mengetahui penilaian status gizi dapat diketahui melalui pengukuran dengan beberapa parameter, selanjutnya hasil pengukuran yang telah didapat akan dibandingkan dengan rujukan. Tujuan dari dilakukannya penilaian status gizi yaitu untuk dapat mengetahui ada tidaknya status gizi yang kurang tepat. Adanya masalah gizi utama seperti Kekurangan Energi Protein (KEP), Gangguan Akibat Kekurangan Yodium, dan Obesitas dan lain-lain menjadikan peran penilaian status gizi sangat penting dilakukan dalam upaya memperbaiki tingkat kesehatan dan resiko kematian yang terkait dengan status gizi pada masyarakat [17].

3 DATA

Data diperoleh melalui pengambilan data di Puskesmas Baqa Samarinda Seberang. Dalam teknik pengambilan data yang digunakan adalah data sekunder dari Puskesmas Baqa Samarinda Seberang. Adapun data terkait yang digunakan adalah data gizi balita mengenai status gizi balita, umur, berat badan dan tinggi badan tahun 2022. Teknik pengambilan data dalam penelitian ini adalah pengambilan data sekunder.

Tabel 3.1 Variabel Penelitian

Variabel Penelitian	Keterangan	Jenis Data
Status Gizi Balita (Y)	Status gizi balita dikatakan baik bila berat badan menurut umur yang dihitung menurut skor Z nilainya lebih dari dari -2 dan gizi buruk bila skor Z kurang dari -3	Kategorik
Umur (X_1)	Umur atau usia balita yang telah terdaftar di Puskesmas pada tahun 2022	Numerik
Berat Badan (X_2)	Berat badan balita yang sesuai dengan pengukuran pada tahun 2022	Numerik
Tinggi Badan (X_3)	Tinggi badan balita yang sesuai dengan pengukuran pada tahun 2022	Numerik

Teknik analisis data yang dilakukan pada penelitian ini yaitu mengklasifikasikan status gizi balita menggunakan metode K-NN. Penelitian ini menggunakan bantuan *software* komputer yaitu *software* R. Adapun langkah-langkah dalam teknik analisis data pada penelitian ini adalah sebagai berikut:

1. Analisis statistika deskriptif
2. Randomisasi data
3. Standardisasi data
4. Menentukan data *training* dan data *testing*
5. Klasifikasi dengan metode *K-Neirest Neighbor*
   a. Menentukan nilai parameter K, di mana pada penelitian ini akan digunakan parameter $K = 1$, $K = 3$, $K = 5$, $K = 7$ dan $K = 9$.
   b. Menghitung jarak Euclid objek terhadap data *training* yang diberikan berdasarkan dengan persamaan (2.1).
   c. Mengurutkan data yang mempunyai jarak terkecil hingga terbesar.

d. Menentukan hasil klasifikasi dengan menggunakan kategori *nearest neighbor* yang paling banyak.
e. Mengevaluasi hasil klasifikasi dengan menggunakan matriks konfusi dan perhitungan dengan akurasi serta *Press's Q*.

6. Mengevaluasi Hasil Klasifikasi

4 HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini merupakan data status gizi balita yang bersumber dari Puskesmas Baqa Samarinda Seberang Tahun 2022. Data tersebut akan dianalisis menggunakan metode klasifikasi K-NN. Adapun variabel terikat yang digunakan adalah status gizi balita dengan kategori gizi buruk, gizi baik, gizi lebih, resiko gizi lebih, gizi kurang dan obesitas. Sedangkan, untuk variabel bebas yang digunakan adalah umur, tinggi badan dan berat badan.

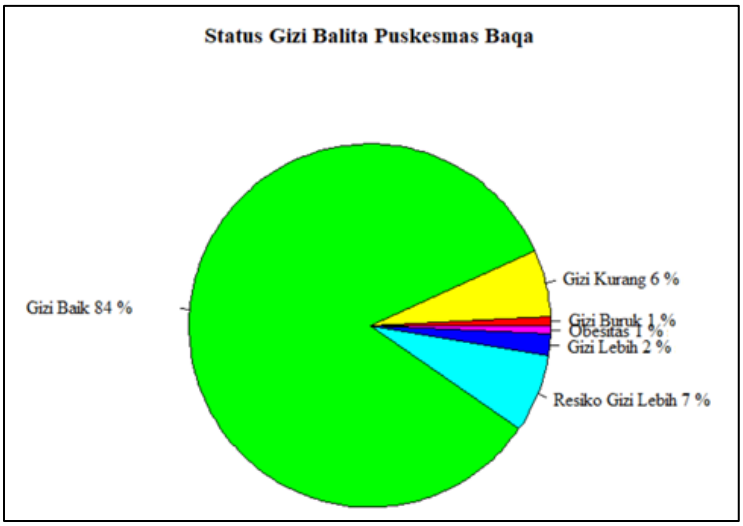
4.1 Analisis Statistika Deskriptif

Analisis statistika deskriptif dilakukan untuk mengetahui gambaran umum data status gizi balita di Puskesmas Baqa Samarinda Seberang. Adapun hasil analisis statistika deskriptif dapat dilihat pada Tabel 4.1 sebagai berikut.

Tabel 4.1 Hasil Analisis Statistika Deskriptif

Variabel	***Minimum***	***Maximum***	**Rata - rata**	**Deviasi Standar**
Umur (X1)	8	69	39,34	15,88
Berat Badan (X2)	2,6	26,3	11,53	3,34
Tinggi Badan (X3)	46	116	86,20	12,75

Pada Tabel 4.1 menunjukkan bahwa pada variabel umur (X1) dengan umur termuda atau terkecil pada umur 8 bulan dengan umur tertua atau terbesar pada umur 69 bulan. Dan rata-rata umur pada balita yaitu 39,34 bulan dengan deviasi standar sebesar 15,58. Lalu, pada variabel berat badan (X2) dengan ukuran berat badan terkecil sebesar 2,6 kg serta ukuran berat badan terbesar sebesar 26,3 kg. Dan rata-rata berat badan balita sebesar 11,53 kg dengan deviasi standar sebesar 3,34. Selanjutnya, pada variabel tinggi badan (X3) dengan ukuran tinggi badan terkecil sebesar 46 cm serta ukuran tinggi badan terbesar atau tertinggi sebesar 116 cm. Dan rata-rata tinggi badan balita sebesar 86,20 cm dengan deviasi standar 12,75. Sedangkan, gambaran umum pada analisis statistika deskriptif mengenai status gizi balita dengan status gizi buruk, gizi kurang, gizi baik, resiko gizi lebih, gizi lebih dan obesitas dapat dilihat pada Gambar 4.1.



Gambar 4.1 Persentase Status Gizi Balita di Puskesmas Baqa Samarinda Seberang

Pada Gambar 4.1 menunjukkan persentase status gizi balita di Puskesmas Baqa Samarinda Seberang Tahun 2022. Dari 531 data balita dapat dilihat bahwa presentase sebesar 1% atau sebanyak 4 orang balita berstatus gizi buruk, persentase sebesar 6% atau sebanyak 32 orang balita berstatus gizi kurang, persentase sebesar 84% atau sebanyak 444 orang balita berstatus gizi baik, persentase sebesar 7% atau sebanyak 37 orang balita berstatus resiko gizi lebih, persentase sebesar 2% atau sebanyak 10 orang balita berstatus gizi lebih dan persentase sebesar 1% atau sebanyak 4 orang balita berstatus obesitas. Dari data tersebut dapat diketahui bahwa balita dengan status gizi baik memiliki persentase yang lebih besar dibandingkan dengan status-status gizi balita lainnya.

4.2 Randomisasi Data

Randomisasi data perlu dilakukan agar semua data memiliki peluang yang sama berkesempatan untuk menjadi data *training* dan data *testing*. Randomisasi data dilakukan dengan menggunakan *software* R. Data hasil randomisasi dapat dilihat pada Tabel 4.2.

Tabel 4.2 Hasil Randomisasi Data

Sampel	Umur (Bulan)	Berat Badan(kg)	Tinggi Badan(cm)	Status Gizi Balita
495	29	10,1	82	Gizi Baik
408	63	16,6	105	Gizi Baik
338	63	17,9	108	Gizi Baik
361	63	16	105	Gizi Baik
454	41	9,9	84	Gizi Baik
⋮	⋮	⋮	⋮	⋮
17	27	9,1	80	Gizi Baik

4.3 Standardisasi Data

Setelah melakukan randomisasi pada data gizi balita di Puskesmas Baqa Samarinda Seberang, selanjutnya akan dilakukan tahap standarisasi data karena terdapat perbedaan ukuran satuan yang cukup besar antar variabel-variabel data dan dapat menyebabkan perhitungan jarak pada analisis klasifikasi menjadi tak valid. Adapun contoh perhitungan standardisasi data menggunakan data training pertama adalah sebagai berikut:

$$Z_{11} = \frac{X_{11} - \bar{X}_1}{S_1} = \frac{29 - 39{,}34}{15{,}86} = -0{,}65164$$
$$Z_{12} = \frac{X_{12} - \bar{X}_2}{S_2} = \frac{10{,}1 - 11{,}53}{3{,}34} = -0{,}42799$$
$$Z_{13} = \frac{X_{13} - \bar{X}_3}{S_3} = \frac{82 - 86{,}2}{12{,}74} = -0{,}32966$$

Selanjutnya, perhitungan standardisasi data dilanjutkan hingga data *training* ke 531. Adapun hasil dari standardisasi data dapat dilihat pada Tabel 4.3 sebagai berikut

Tabel 4.3 Hasil Standarisasi Data

Sampel	Umur (Bulan)	Berat Badan(kg)	Tinggi Badan(cm)	Status Gizi Balita
495	-0,65164	-0,42799	-0,32966	Gizi Baik
408	1,49108	1,51743	1,475603	Gizi Baik
338	1,49108	1,90652	1,711072	Gizi Baik
361	1,49108	1,33785	1,475603	Gizi Baik
454	0,10462	-0,48785	-0,17268	Gizi Baik
⋮	⋮	⋮	⋮	⋮
17	-0,77768	-0,72729	-0,48664	Gizi Baik

4.4 Menentukan data *training* dan data *testing*

Dengan proporsi pembagian 90% data *training* dan 10% data *testing.* Adapun perhitungan pembagian data *training* serta data *testing* dengan menggunakan Persamaan (2.3) dan Persamaan (2.4) sebagai berikut:

Jumlah data *training* = Proporsi data *training* $\times N$
$= 90\% \times 531$
$= 478$
Jumlah data *testing* $= N -$ Jumlah data $training$
$= 531 - 478$
$= 53$

Berdasarkan perhitungan diperoleh hasil bahwa dari 531 data hasil randomisasi gizi balita di Puskesmas Baqa Samarinda Seberang untuk masing-

masing variabel penelitian diperoleh hasil yaitu data *training* proporsi 90% sebanyak 478 data dan data *testing* proporsi 10% sebanyak 53 data.

4.5 Klasifikasi dengan metode *K-Neirest Neighbor*

Pada proses klasifikasi menggunakan metode K-NN terdapat tiga tahapan, yaitu menentukan nilai parameter *K*, menghitung jarak Euclid antara data *training* dengan data *testing*, kemudian menentukan *ranking* dari hasil perhitungan jarak Euclid. Adapun tahapan-tahapan dalam proses pengklasifikasian dengan metode K-NN sebagai berikut:

a. Menentukan nilai parameter K
   Klasifikasi status gizi balita menggunakan metode K-NN dilakukan dengan menentukan nilai parameter *K* terlebih dahulu. Adapun nilai *K* yang digunakan yaitu $K = 1,\ K = 3,\ K = 5, K = 7$ dan $K = 9$. Pada kali ini akan diberikan contoh perhitungan menggunakan parameter $K = 1$.

b. Menghitung jarak Euclid
   Dilakukan perhitungan jarak Euclid pada objek antara data *testing* dan data *training* berdasarkan dengan Persamaan (2.1). Adapun perhitungan jarak euclid antara data *testing* pertama dan data *training* pertama hingga data training ke - 478 sebagai berikut:

$$d(x_1 - x_1^*) = \sqrt{(x_{11} - x_{11}^*)^2 + (x_{12} - x_{12}^*)^2 + (x_{13} - x_{13}^*)^2}$$

$$= \sqrt{\begin{matrix}\left((-0{,}65192) - (0{,}29380)\right)^2 + \left((-0{,}42799) - (-0{,}21848)\right)^2 \\ +\left((-0{,}32965) - 0{,}09418\right)^2\end{matrix}}$$

$$= 0{,}99686$$

$$d(x_2 - x_1^*) = \sqrt{(x_{21} - x_{21}^*)^2 + (x_{22} - x_{22}^*)^2 + (x_{23} - x_{23}^*)^2}$$

$$= \sqrt{\begin{matrix}\left((-1{,}49108) - (0{,}29380)\right)^2 + \left((1{,}51743) - (-0{,}21848)\right)^2 \\ +\left((1{,}4756) - 0{,}09418\right)^2\end{matrix}}$$

$$= 2{,}62919$$

$$\vdots$$

$$d(x_{478} - x_1^*) = \sqrt{(x_{4781} - x_{4781}^*)^2 + (x_{4782} - x_{4782}^*)^2 + (x_{4783} - x_{4783}^*)^2}$$

$$= \sqrt{\begin{matrix}\left((-0{,}77768) - (0{,}29380)\right)^2 + \left((-0.39806) - (-0{,}21848)\right)^2 \\ +\left((-0{,}64361) - 0{,}09418\right)^2\end{matrix}}$$

$$= 1{,}21775$$

c. Menentukan ranking dari hasil perhitungan jarak
   Setelah melakukan perhitungan jarak menggunakan Persamaan (2.1), kemudian dilanjutkan dengan mengurutkan (*ranking*) hasil perhitungan

jarak untuk menemukan urutan terkecil hingga terbesar. Adapun *ranking* jarak Euclid dapat dilihat pada Tabel 4.4 sebagai berikut.

Tabel 4.4 Ranking Hasil Jarak Euclid antara Data *Testing* Pertama dan Data *Training*

Rank	**Data *Training***		**Data *Testing* Pertama**	**Batas K-NN**
	Sampel	**Klasifikasi**	**d**	
1	212	Gizi Baik	0,11713	1-NN
2	465	Gizi Baik	0,16014	-
3	166	Gizi Baik	0,1915	-
4	227	Gizi Baik	0,21235	-
5	214	Gizi Baik	0,22481	-
6	184	Gizi Baik	0,23721	-
7	180	Gizi Baik	0,25974	-
8	108	Gizi Baik	0,26781	-
9	33	Gizi Baik	0,27061	-
⋮	⋮	⋮	⋮	⋮
478	62	Obesitas	5,19835	-

Berdasarkan Tabel 4.4 dapat diketahui bahwa dengan menggunakan 1-NN, pada data testing pertama yaitu balita dengan umur (X1) yaitu 44 bulan, memiliki berat badan (X2) sebesar 10,6 kg dan tinggi badan (X3) sebesar 84 cm, diprediksi berstatus gizi baik.

4.6 Mengevaluasi Hasil Klasifikasi

Pada proses klasifikasi menggunakan metode K-NN jumlah obyek yang tepat dan salah diklasifikasikan untuk masing-masing kelompok dapat dilihat pada Tabel 4.5 Tanda (*) pada angka-angka menyatakan jumlah obyek kelompok tertentu yang benar diklasifikasikan dengan menggunakan metode K-NN.

Tabel 4.5 Hasil Klasifikasi Metode Klasifikasi *K-Neirest Neighbor* dengan K=1

Klasifi-kasi Awal Status Gizi Balita	**Prediksi Klasifikasi Metode K-NN**						**Total**
	Gizi Baik	Gizi Buruk	Gizi Kurang	Gizi Lebih	Obesitas	Resiko Gizi Lebih	
Gizi Baik	44*	0	2	0	0	1	47
Gizi Kurang	1	0	0*	0	0	0	1
Gizi Lebih	2	0	0	0*	0	1	3
Resiko Gizi Lebih	0	0	0	0	0	2*	2
Total	47	0	2	0	0	4	53

Berdasarkan Tabel 4.5 dapat diketahui bahwa status gizi balita di Puskesmas Baqa Samarinda Seberang menggunakan metode K-NN diperoleh hasil yaitu dari 47 balita yang memiliki status gizi baik, terdapat 44 balita yang tepat diklasifikasikan memiliki status gizi baik dan 3 lainnya tidak tepat diklasifikasikan. Sedangkan dari 1 balita yang memiliki status gizi kurang, terdapat 0 balita yang tepat diklasifikasikan memiliki status gizi kurang dan 1 lainnya tidak tepat diklasifikasikan. Selanjutnya, dari 3 balita yang memiliki status gizi lebih terdapat 0 balita yang tepat diklasifikasikan memiliki status gizi lebih dan 3 lainnya tidak tepat diklasifikasikan. Lalu, dari 2 balita yang memiliki status resiko gizi lebih terdapat 2 gizi balita tepat diklasifikasikan. Sehingga berdasarkan Persamaan (2.5), maka diperoleh nilai akurasi sebagai berikut:

$$\begin{aligned}\text{Akurasi} &= \frac{44+0+0+2}{53} \times 100\% \\ &= \frac{46}{53} \times 100\% \\ &= 86{,}79\%\end{aligned}$$

Jadi, nilai akurasi pada hasil klasifikasi menggunakan 1-NN adalah sebesar 86,79%. Dengan cara yag sama peneliti juga melakukan prediksi menggunakan K = 3, 5, 7, dan 9. Untuk menentukan nilai K optimal, dapat dilakukan dengan melihat nilai akurasi dengan bantuan *software* R. Adapun nilai akurasi untuk seluruh nilai K yang dicobakan dapat dilihat pada Tabel 4.6 sebagai berikut.

Tabel 4.6 Nilai Akurasi Pada Metode K-NN untuk Seluruh Nilai K

Nilai K	Akurasi
1	86,79%
3	90,57%
5	90,57%
7	90,57%
9	90,57%

Selanjutnya, untuk mengevaluasi hasil ketepatan hasil prediksi (klasifikasi) maka digunakan nilai *Press's Q*. Perhitungan nilai *Press's Q* dengan menggunakan matriks konfusi pada Tabel 4.5 dan Persamaan (2.6) sehingga didapatkan hasil sebagai berikut.

$$\begin{aligned}Press's\ Q &= \frac{[53-((46)6)]^2}{53(6-1)} \\ &= 187{,}65\end{aligned}$$

Dengan derajat bebas bernilai satu *s*erta tingkat kepercayaan $\alpha = 0{,}05$, dan nilai X^2 tabel = 3,841 < *Press's Q* = 187,65. Maka dapat disimpulkan bahwa pengklafikasian menggunakan metode K-NN akurat.

5 KESIMPULAN

Berdasarkan dengan penelitian yang telah dilakukan maka dapat disimpulkan bahwa

1. Hasil klasifikasi menggunakan metode K-NN dengan proporsi data *training* 90% : data *testing* 10% dengan K =1 menghasilkan akurasi sebesar 86,79%, dengan K=3, 5, 7, dan 9 menghasilkan akurasi sebesar 90,57%. Dimana K=3,5,7 dan 9 merupakan nilai akurasi terbesar dibandingkan dengan nilai akurasi K=1.
2. Hasil klasifikasi menggunakan metode K-NN memiliki nilai *Press's Q* sebesar 187,65 dimana nilai tersebut membuktikan bahwa pengklasifikasian menggunakan metode K-NN akurat untuk digunakan.

Sedangkan saran yang dapat diberikan untuk penelitian selanjutnya yaitu:

1. Penelitian selanjutnya dapat dikembangkan dengan menggunakan data pada variabel bebas lebih dari tiga variabel atau menggunakan variabel pendukung lainnya.
2. Penelitian selanjutnya dapat dikembangkan menggunakan perhitungan evaluasi ketepatan klasifikasi lain seperti menggunakan perhitungan *Recall* dan *Precision*, serta *Hold-Out*.

DAFTAR PUSTAKA

[1] Han, J., Kamber, M., & Pei, J. (2012) . *Data Mining: Concept and Techniques, Third Edition*. Waltham: Morgan Kaufmann Publishers.

[2] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Partical Machine Learning Tools and Techniques, Third Edition*. Burlington: Morgan Kaufmann Publishers.

[3] Yustanti, W. (2012). Algoritma K-Nearest Neighbor untuk Memprediksi Harga Jual Tanah. *Jurnal Matematika, Statistika, Matematika:JMSK. 9*(1), 57-68.

[4] Rahmaulidyah, F.N., Hayati, M.N., & Goejantoro, R. (2021). Perbandingan Metode Klasifikasi *naïve* Bayes dan *K-Neirest Neighbor* Pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Samarinda Ulu. *Jurnal Eksponensial, 12(2*), 161-164.

[5] Mustaghfiroh, L., Ariani, M. H., & Bijanto. (2022) . Klasifikasi Pasien Covid-19 di Indonesia menggunakan Metode *K-Nearest Neighbor*. *Jurnal AMRI, 1*(1), 16-21.

[6] Saeroni, A., Hayati, M.N., & Goejantoro, R. (2020). Klasifikasi Tingkat Kelancaran Nasabah Dalam Membayar Premi dengan Menggunakan Metode *K-Nearest Neighbor* dan Analisis Diskriminan Fisher (Studi kasus: Data Nasabah PT. Prudential Life Samarinda Tahun 2019). *Jurnal Statistika Universitas Muhammadiyah Semarang, 8(2),* 88-94.

[7] Sibagariang, E. (2010) . *Gizi Dalam Kesehatan Reproduksi*. Jakarta: Trans Info Media.

[8] Kementrian Kesehatan RI. (2015). *Situasi Kesehatan Anak Balita di Indonesia*. Jakarta: Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia

[9] Turban, E., Aronson, J. E., & Liang, T. P. (2005). *Decision Support Systems and Intelligent Systems*. Yogyakarta: Andi Offset.

[10] Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons Inc.
[11] Prasetyo, E. (2014) . *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Andi Offset.
[12] Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Boston: Pearson Education.
[13] Simamora, B. (2005). Analisis Multivariat Pemasaran. Jakarta: PT. Gramedia Pustaka Utama.
[14] Prasetyo, E. (2012) . *Data Mining: Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.
[15] Hair, J., Black, W., Babin, B., Anderson, R. & Tatham, R. (2006) . *Multivariate Data Analysis 5th Edition* . New Jersey: Pearson Prentice Hall.
[16] Supariasa, I. D. N., Bakri, B., & Fajar, I. (2002) . *Penilaian Status Gizi*. Jakarta: Buku Kedokteran EGC.
[17] Harjatmo, T. P., Par'i, H. M., & Wiyono, S. (2017) . *Buku Ajar Penilaian Status Gizi*. Jakarta: Kementrian Kesehatan Republik Indonesia.