Penentuan Model Terbaik pada Metode *Naive Bayes Classifier* dalam Menentukan Status Gizi Balita dengan Mempertimbangkan Independensi Parameter

Apriliya Fitri Cahyanti
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami 36A Kentingan
Surakarta

apriliya.fc@student.uns.ac.id

Ristu Saptono
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami 36A Kentingan
Surakarta

ristu.saptono@staff.uns.ac.id

Sari Widya Sihwi
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami 36A Kentingan
Surakarta

sari.widya.sihwi@gmail.com

ABSTRAK

Untuk proses klasifikasi dalam studi kasus penilaian status gizi balita menggunakan metode *Naïve Bayes Classifier*, asumsi independensi antar parameter perlu diperhitungkan. Independensi antar parameter dilihat dari korelasi antar parameter yang digunakan. Artikel ini membahas mengenai uji korelasi antar parameter dalam studi kasus penilaian status gizi menggunakan metode *Cosine Similarity*. Kemudian hasil uji korelasi tersebut dijadikan prosedur penentuan model dalam metode *Naive Bayes Classifier*. Sehingga dapat diketahui model yang paling baik dalam penilaian status gizi menggunakan metode *Naive Bayes Classifier*. Penentuan model terbaik dilihat dari akurasi, kesederhanaan, waktu, dan akuisisi data pada model. Pada skenario data 60%:40%, model terbaik ditunjukkan oleh model yang terdiri dari parameter berat, bmi, dan umur, dengan akurasi sebesar 94.4%. Sedangkan pada skenario data 80%:20% model terbaik ditunjukkan pada model yang terdiri dari parameter berat, bmi, tinggi, umur, dan jenis kelamin, dengan akurasi 94,8%. Penelitian ini menunjukkan bahwa kolerasi parameter mempengaruhi hasil klasifikasi. Penggunaan parameter independen belum tentu menghasilkan akurasi yang maksimal. Bahkan, model terbaik yang dipilih terdiri dari parameter dependen.

Kata Kunci *— Cosine Similarity, Naive Bayes Classifier*, Penilaian Status Gizi.

1. PENDAHULUAN

Naïve Bayes Classifier merupakan metode *classifier* yang berdasarkan probabilitas dan teorema Bayesian dengan asumsi keindependenan atribut [1]. Asumsi independensi atribut akan menghilangkan kebutuhan banyaknya jumlah data latih dari seluruh atribut yang dibutuhkan untuk mengklasifikasi suatu data. Data latih untuk Teorema Bayes membutuhkan paling tidak perkalian kartesius dari seluruh kelompok atribut yang mungkin, sehingga semakin sedikit atribut yang digunakan, akan mengurangi data latih yang dibutuhkan. Padahal, pada kenyataannya asumsi atribut independen pada *Naïve Bayes Classifier* sering dilanggar [2]. Hal ini disebabkan karena asumsi keindependenan atribut dalam dunia nyata hampir tidak pernah terjadi [3].

Independensi parameter sendiri ditentukan berdasarkan nilai korelasi atau kemiripan parameter yang digunakan pada klasifikasi. Dengan menggunakan metode *Cosine Similarity* dapat diketahui similaritas antar parameter, dan korelasi antar perameter tersebut, sehingga dapat menentukan independensi antar parameter. Pada penelitian yang dilakukan oleh Domingos dan Pazzani, ternyata menghasilkan *Naive Bayes* memiliki performa yang baik meskipun di dalamnya terdiri dari atribut yang dependen. [2]

Dalam penelitian ini penulis menggunakan studi kasus mengenai penilaian status gizi. Penilaian status gizi sendiri menggunakan parameter berupa ukuran tubuh seperti berat badan, tinggi badan, lingkar lengan atas dan tebal lemak di bawah kulit [4]. Penilaian status gizi menggunakan parameter yang dibuat menjadi kategorikal, sehingga masing – masing parameter saling berkaitan. Penelitian ini membahas mengenai korelasi antara parameter dan hasil klasifikasi, dan uji independensi antar parameter menggunakan metode *Cosine Similarity*. Penulis ingin mengeksplore lebih lanjut bagaimana keoptimalan *Naive Bayes Classifier* pada studi kasus penilaian status gizi dengan mempertimbangkan independensi parameter, dengan pemodelan data menggunakan *Backward Feature Selection.* Pemodelan data *Backward Feature Selection* dilakukan dengan menghapus fitur, yang apabila fitur tersebut dihapus dapat meningkatkan akurasi sistem. Selanjutnya, dipilih model yang paling baik dengan mempertimbangkan akurasi, kesederhanaan model, waktu eksekusi, dan akuisisi data model.

2. LANDASAN TEORI

2.1. *Naïve Bayes Classifier*

*Naive Bayes*ian *Classifier* merupakan klasifikasi dengan model statistik untuk menghitung peluang dari suatu kelas yang memiliki masing – masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal. Pada metode ini semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain [5].

Dasar dari teorema Naïve Bayes Classifier yang dipakai dalam pemrograman adalah rumus Bayes sebagai berikut [8].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Dimana

P(H|X) = probabilitas posterior H di dalam X
P(X|H) = Probabilitas posterior X di dalam H
P(H) = Probabilitas prior dari H
P(X) = Probabilitas prior dari X

Dalam sebuah dataset yang besar, pemilihan data *training* secara *random* akan menyebabkan kemungkinan adanya nilai nol dalam model probabilitas. Nilai nol ini akan menyebabkan *Naive Bayes Classifier* tidak dapat mengklasifikasi sebuah data inputan. Oleh karena itu diperlukan suatu metode smoothing yang dapat menghindari adanya nilai nol dalam model probabilitas. *Laplacian Smoothing* merupakan metode smoothing yang biasa digunakan dalam *Naive Bayes* Classifier. *Laplacian Smoothing* biasa dikenal dengan nama *add one smoothing*, karena dalam perhitungannya, setiap variabel pada masing – masing parameter ditambahkan 1. Persamaan *Laplace Smoothing* dituliskan pada persamaan (2).

$$P(x_k|C) = \frac{P(x_k|C)+1}{P(C)+|V|} \quad (2)$$

Dimana

$P(x_k|C)$= probabilitas tiap atribut dari xk
$P(C)$ = total jumlah probabilitas dalam xk
|V| = jumlah kemungkinan nilai dari xk.

2.2. *Backward Feature Selection*

Seleksi fitur atau *feature selection* digunakan untuk memilih fitur sesuai kriteria yang telah ditentukan. Pemodelan data Backward Feature Selection dilakukan dengan menghapus fitur, yang apabila fitur tersebut dihapus dapat meningkatkan akurasi sistem. Selain itu, jika dengan menghapus fitur tersebut berdampak pada penurunan performa, maka fitur tersebut dianggap sangat diperlukan dalam proses klasifikasi [6]. Pada Backward Feature Selection, dimulai dengan menggunakan semua parameter yang ada. Selanjutnya menghapus satu per satu parameter, yang pada setiap penghapusan dapat menurunkan tingkat kesalahan sistem [7].

Iterasi	*Feature Set*	*Score*	Hapus Fitur
Iterasi 0	[A B C D]	0.41	
Iterasi 1	[A B C]	0.24	C
	[A B D]	0.43	
	[A C D]	0.36	
	[B C D]	0.19	
Iterasi 2	[A B]	0.45	B
	[A D]	0.58	
	[B D]	0.29	
Iterasi 3	[A]	0.68	D
	[D]	0.47	

Gambar 1. Simulasi *Backward Feature Selection*

Berdasarkan simulasi diatas, pembentukan fitur set dimulai dengan fitur set yang lengkap. Selanjutnya dilakukan penghapusan satu persatu fitur, hingga ditentukan fitur set yang paling baik digunakan adalah A pada iterasi 3 dengan nilai 0.68. Penghapusan fitur B, C, dan D pada iterasi 1 dan iterasi 2 dapat meningkatkan nilai performa.

2.3. *Cosine Similarity*

Metode *Cosine* digunakan untuk menghitung nilai cosinus sudut antara dua *vector* dan mengukur kemiripan antar dua dokumen [8]. Untuk membandingkan dua parameter A dan B, maka perhitungan *similarity* dapat dihitung dengan persamaan (3).

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} Ai \times Bi}{\sqrt{\sum_{i=1}^{n}(Ai)^2} \times \sqrt{\sum_{i=1}^{n}(Bi)^2}} \quad (3)$$

Dimana

A = bobot data yang dibandingkan
B = bobot data pembanding
||A|| = panjang data yang dibandingkan
||B|| = panjang data pembanding

Cosine Similarity berpusat x dan y, dan dibatasi antara -1 dan 1. Namun, nilai cosinus dibatasi 0 dan 1 jika x dan y bernilai positif. Menurut O'Connor, *Cosine Similarity* dapat menunjukkan korelasi. Pada umumnya *Cosine Similarity* dibahas dalam hal sudut vektor, tetapi dapat dianggap sebagai korelasi, jika vektor yang dibandingkan merupakan suatu data yang berpasangan [9]. Dua parameter dengan nilai similaritas besar menunjukkan parameter yang saling berkorelasi.

2.4. *Korelasi*

Informasi korelasi diketahui berdasarkan pada nilai kemiripan [10]. Semakin besar nilai kemiripan dari kedua objek, maka menunjukkan hubungan atau korelasi parameter yang kuat. Sedangkan parameter yang memiliki korelasi tinggi menunjukkan parameter yang dependen. Kekuatan korelasi antar parameter ditunjukkan oleh koefesien korelasi.

2.5. *Stratified Random Sampling* atau *Holdout*

Metode *Holdout* merupakan metode yang menyediakan sejumlah data untuk digunakan sebagai data *testing* dan sisanya sebagai data *training* [11]. Dalam penelitian ini data dibagi menjadi dua kelompok, yaitu data *training* dan data pengujian yang dibagi secara acak. Data pelatihan digunakan untuk memperoleh model, sedangkan data pengujian digunakan untuk mengestimasi akurasi.

Dalam pengacakan data *training* dan *testing* ini ada kemungkinan akan menghasilkan data yang tidak proporsional dalam tiap klasifikasi. Misalnya satu klasifikasi data lebih dominan jika dibandingkan klasifikasi yang lain. Oleh karena itu digunakan metode stratified random sampling untuk pengacakan data untuk menghasilkan data *training* dan *testing* yang proposional.

2.6. Penilaian Status Gizi

Antropometri sebagai indikator status gizi dapat dilakukan dengan mengukur beberapa parameter. Parameter adalah ukuran tunggal dari tubuh manusia yaitu umur, berat badan, tinggi badan, lingkar lengan atas, lingkar dada, lingkar panggul, dan tebal lemak di bawah kulit [4]. Standar antropometri pada balita berbeda untuk tiap jenis kelamin, baik laki – laki maupun perempuan. Berdasarkan berat dan tinggi anak, dapat diketahui *Body Mass Index* (BMI) yang juga dapat menentukan nilai status gizi anak tersebut.

Penggunaan *Body Mass Index* (BMI) pada umumnya hanya berlaku untuk orang dewasa. Pada anak – anak pengukuran BMI sangat terkait dengan umurnya, karena dengan perubahan umur terjadi perubahan komposisi tubuh. Karena itu, pada anak – anak digunakan indeks BMI menurut umur (BMI/U). BMI dapat dihitung dengan rumus:

$$BMI = \frac{berat\ badan\ (kg)}{tinggi\ badan^2(m)} \quad (4)$$

Selain itu, menurut [4], faktor yang mempengaruhi keadaan gizi yaitu konsumsi makanan dan tingkat kesehatan. Sedangkan konsumsi makanan sendiri dipengaruhi oleh pendapatan orang tua, makanan, dan tersedianya bahan makanan. Terdapat hubungan antara tingkat pendidikan dan pendapatan keluarga terhadap status gizi balita [13]. Pendapatan keluarga sendiri menyebabkan keluarga masuk ke kategori keluarga miskin (gakin) atau tidak).

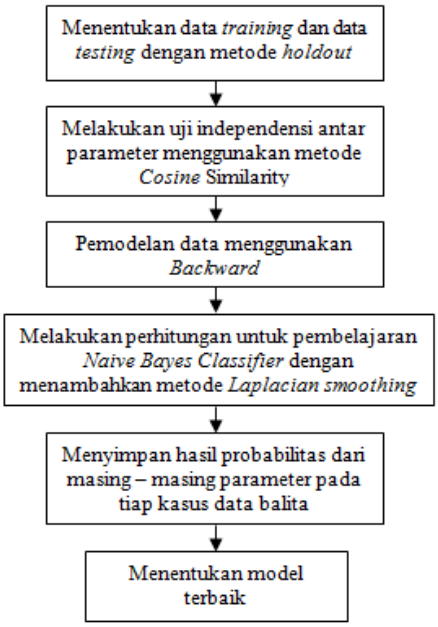
3. METODOLOGI PENELITIAN

3.1. Pengumpulan Data

Dataset yang digunakan merupakan data posyandu pada bulan September 2013 yang diambil dari puskesmas kelurahan Cangkrep, Kecamatan Purworejo, Kabupaten Purworejo. Jumlah seluruh dataset yaitu 250 data.

3.2. Pemodelan Data

Pemodelan data dilakukan untuk mengetahui alur bagaimana proses – proses dan metode berjalan sebelum diimplementasikan ke dalam sebuah aplikasi dengan data yang ada. Berikut tahap pemodelan data untuk penentuan status gizi menggunakan *Naïve Bayes Classifier*.



Gambar 2. Alur Pemodelan Data

Dalam proses pengkategorian parameter menurut Kementerian Kesehatan, beberapa kategori parameter dipengaruhi oleh parameter yang lain [12]. Kategori berat badan ditentukan oleh umur dalam bulan, jenis kelamin, dan berat dalam kg. Pada kategori tinggi badan, ditentukan oleh umur dalam bulan, jenis kelamin, dan tinggi dalam cm. Sedangkan untuk kategori BMI (*Body Mass Index*) ditentukan oleh umur dalam bulan, jenis kelamin, berat dalam kg, dan tinggi dalam cm. Untuk label pengkategorian dapat dilihat pada tabel 3.1 berikut ini.

Tabel 3.1 Diskritisasi Parameter

No	Parameter	Kategori	Label
1	Status gizi	Gizi buruk	0
		Gizi kurang	1
		Gizi baik	2
		Gizi Lebih	3
2	Jenis Kelamin	Perempuan	0
		Laki-Laki	1
3	Status ekonomi keluarga (Keluarga Miskin/Gakin)	Ya	0
		Tidak	1
4	Berat Badan	Sangat Kurang	0
		Kurang	1
		Baik	2
		Lebih	3
5	Tinggi Badan	Sangat Pendek	0
		Pendek	1
		Normal	2
		Tinggi	3
6	*Body Mass Index* (BMI)	Sangat Kurus	0
		Kurus	1
		Normal	2
		Gemuk	3
7	Umur (bulan)	0 sd. 60	

Penelitian ini mengasumsikan data yang digunakan adalah bersifat kategorikal, sehingga untuk data yang bertipe kontinu harus melalui *preprocessing* diskritisasi untuk menghasilkan data yang bersifat kategorikal. Angka binary yang digunakan untuk label kategorikal ini dimulai dari 0. Untuk parameter umur sendiri, tidak dilakukan diskritisasi, karena nilai umur yaitu 0 sampai 60 sudah dianggap bersifat kategorikal. Untuk parameter yang lain, yaitu jenis kelamin, status ekonomi, berat badan, tinggi badan, dan BMI, pengkategorian berdasarkan pada standar antropometri yang ditetapkan oleh Kementrian Kesehatan RI.

3.2.1 Pengacakan Data menggunakan Metode *Holdout*

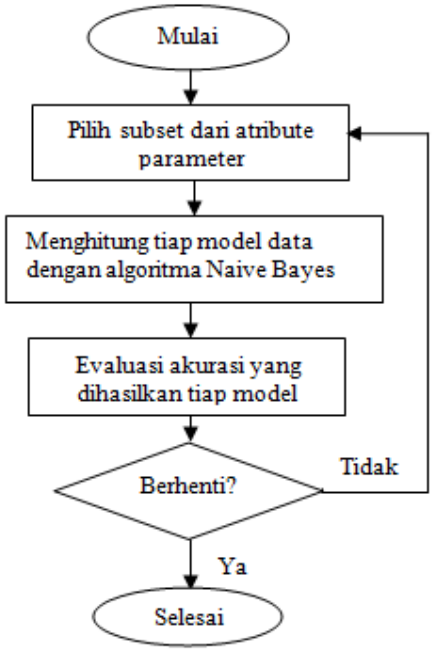
Untuk pembagian data *training* dan data *testing* dilakukan secara acak dengan persentase 60% untuk data *training* dan 40% untuk data *testing* , serta 80% untuk data *training* dan 20% untuk data *testing* . Untuk masing – masing skenario, dilakukan 5 kali percobaan.

3.2.2 Uji Korelasi Parameter dengan *Cosine Similarity*

Uji korelasi dilakukan pada tiap parameter yang digunakan (jenis kelamin, umur, berat, tinggi, dan gakin) terhadap parameter status gizi. Dari hasil uji korelasi tersebut, parameter *independent* yang memiliki nilai *similarity* terbesar berarti memiliki pengaruh paling tinggi terhadap parameter *dependent* (status gizi). Selanjutnya, melakukan uji korelasi antar parameter menggunakan *Cosine* untuk menentukan independensi antar parameter.

3.2.3 Pembentukan Model Data

Uji independensi parameter menentukan model data yang akan diujikan menggunakan metode *Naïve Bayes Classifier*. Model data ini dibentuk berdasarkan urutan similaritas parameter yang tersedia. Prosedur penentuan model dilakukan secara *Backward Feature Selection* dengan menghilangkan satu persatu parameter mulai dari parameter yang memiliki similaritas terkecil.



Gambar 3. Alur Pembentukan Model Data

3.3. Pengembangan Sistem Aplikasi

Pada penelitian ini dilakukan pengembangan suatu aplikasi untuk menentukan model yang paling baik dalam penilaian status gizi balita dengan menggunakan bahasa pemrograman PHP dan database MySql.

3.4. Pengujian dan Analisis Hasil

Untuk mengetahui kinerja dari sistem yang dibuat, dilakukan pengujian dengan membandingkan hasil klasifikasi menggunakan software WEKA dengan hasil klasifikasi pada sistem. Hasil yang dibandingkan adalah hasil akurasi dari kedua sistem. Jika hasil model terbaik dari WEKA dan sistem sama, maka dapat dikatakan hasil sistem benar.

Dalam percobaan, dataset yang ada dibagi menjadi data *training* dan data *testing* . Terdapat dua prosedur dalam pembagian data *training* dan *testing* dalam persentase data yang berbeda. Dengan adanya prosedur yang berbeda tersebut dapat dilihat performa metode yang diuji berdasarkan masing – masing prosedur.

Tabel 3.3 Skenario Pengujian Data

Percobaan	Persentase Data	
	Training	*Testing*
Percobaan 1	60%	40%
Percobaan 2	80%	20%

Untuk mengukur tingkat akurasi dari metode *Naive Bayes Classifier* dihitung menggunakan persamaan (5).

$$Akurasi = \frac{jumlah\ prediksi\ benar}{Jumlah\ Data} \times 100\% \qquad (5)$$

Selanjutnya parameter lain yang diukur adalah waktu eksekusi dari sistem dalam satuan detik. Selain itu, untuk menentukan model terbaik dilihat akuisisi parameter yang terdapat pada model. Akuisisi parameter yaitu penentuan parameter lapangan yang cocok dengan daerah survey.

4. HASIL DAN PEMBAHASAN

4.1. Deskripsi Data

Data yang digunakan dalam penelitian ini sebanyak 250 data balita. Data tersebut diambil dari data posyandu Puskesmas Kecamatan Purworejo. Data balita yang disimpan terdiri dari parameter umur, jenis kelamin, berat, tinggi, bmi, dan gakin.

4.2. Pemodelan Data

4.1.1. Uji Korelasi Parameter dengan *Cosine Similarity*

Dari beberapa parameter yang digunakan, terlebih dahulu dilakukan uji korelasi dengan menghitung nilai *similarity* dari tiap parameter terhadap status gizi. Nilai *similarity* yang besar berarti nilai independensinya kecil. Uji korelasi ini dilakukan menggunakan metode *Cosine Similarity*, dan diurutkan berdasarkan nilai *similarity* paling besar. Karena *Cosine Similarity* tidak memiliki batas untuk koefisien korelasi, maka penulis memberikan batasan sebesar 0,4 sebagai batas untuk parameter dependen.

Tabel 4.1 Uji Korelasi Status Gizi Dengan Semua Parameter pada Skenario 60%:40%

Parameter dependent	**Parameter independent**	**Similarity**
Status gizi	berat	0.99333
Status gizi	bmi	0.97072
Status gizi	tinggi	0.88979
Status gizi	umur	0.80056
Status gizi	jenis kelamin	0.64215
Status gizi	gakin	0.29101

Berdasarkan hasil uji korelasi tersebut, terlihat bahwa berat memiliki angka *similarity* yang paling tinggi yaitu 0.99333. Hal ini menunjukkan parameter berat memiliki keterkaitan yang tinggi dengan status gizi. Sedangkan parameter gakin memiliki angka similaritas paling kecil dengan nilai 0,29101. Parameter gakin dapat dikatakan sebagai parameter yang independen terhadap status gizi. Selanjutnya dilakukan uji korelasi antar parameter untuk mengetahui hubungan antar parameter.

Tabel 4.2 Uji Korelasi Antar Parameter pada Skenario 60%:40%

Parameter	berat	bmi	tinggi	umur	jenis kelamin	gakin
berat	1	0.970	0.891	0.790	0.638	0.281
bmi	0.970	1	0.862	0.848	0.671	0.329
tinggi	0.891	0.862	1	0.755	0.580	0.329
umur	0.790	0.848	0.755	1	0.619	0.370
jenis kelamin	0.638	0.671	0.580	0.619	1	0.314
gakin	0.281	0.329	0.329	0.370	0.314	1

Hasil uji korelasi antar parameter dapat dilihat pada tabel 4.2. Urutan parameter yang memiliki similaritas paling tinggi dengan parameter berat adalah BMI, tinggi, umur, jenis kelamin, kemudian gakin. Nilai similaritas bmi adalah 0.970, dan nilai similaritas gakin adalah 0.281.

Tabel 4.3 Uji Korelasi Status Gizi Dengan Semua Parameter pada Skenario 80%:20%

Parameter dependent	**Parameter independent**	**Similarity**
Status gizi	berat	0.99321
Status gizi	bmi	0.97226
Status gizi	tinggi	0.88111
Status gizi	umur	0.79924
Status gizi	jenis kelamin	0.65397
Status gizi	gakin	0.27398

Tabel 4.3 diatas menunjukkan hasil uji korelasi status gizi dengan semua parameter yang dilakukan pada skenario 80%:20%. Sama halnya dengan skenario 60%:40%, pada skenario ini juga menunjukkan parameter berat memiliki nilai similaritas tertinggi dengan status gizi, dengan nilai 0.99321. Urutan parameter dari nilai similaritas tertinggi dari kedua skenario juga sama, yaitu berat, bmi, tinggi, umur, jenis kelamin, kemudian gakin.

Tabel 4.4 Uji Korelasi Antar Parameter pada Skenario 80%:20%

Parameter	berat	bmi	tinggi	umur	jenis kelamin	gakin
berat	1	0.971	0.884	0.789	0.651	0.272
bmi	0.971	1	0.861	0.852	0.681	0.298
tinggi	0.884	0.861	1	0.764	0.572	0.296
umur	0.789	0.852	0.764	1	0.607	0.348
jenis kelamin	0.651	0.681	0.572	0.607	1	0.251
gakin	0.272	0.298	0.296	0.348	0.251	1

Pada uji korelasi antar parameter pada skenario 80%:20% juga menghasilkan urutan yang sama dengan uji korelasi pada skenario 60%:40%. Urutan parameter berdasarkan niilai similaritas tertinggi terdapat pada parameter bmi, mempunyai nilai sebesar 0.971, kemudian tinggi, umur, jenis kelamin, dan yang terakhir adalah gakin dengan nilai similaritas sebesar 0.272.

4.1.2. Pembentukan Model Data *Backward*

Uji independensi antar parameter akan membentuk prosedur model data yang akan diujikan menggunakan metode *Naïve Bayes Classifier*. Model ini didapatkan menggunakan metode *backward* dengan mengurangi satu persatu parameter, sehingga hanya tersisa satu parameter. Model data yang dihasilkan berdasarkan proses uji independensi adalah sebanyak 21 model.

Pemodelan data dibagi menjadi enam bagian berdasarkan jumlah parameternya. Pertama terdiri dari enam parameter atau parameter lengkap yaitu model 1. Yang kedua terdiri dari lima parameter yaitu model 2, 3, 4, 5, 6, dan 7. Bagian ketiga terdiri dari empat parameter yaitu model 8, 9, 10, 11, dan 12. Bagian keempat terdiri dari tiga parameter yaitu model 13, 14, 15, dan 16. bagian kelima terdiri dari dua parameter yaitu model 17, 18, dan 19. Sedangkan bagian terakhir terdiri dari satu parameter yaitu model 20 dan 21.

Tabel 4.5 Model Data pada Skenario 60%:40%

Model	Parameter yang digunakan	Akurasi
1	berat, bmi, tinggi, umur, jenis kelamin, gakin	96.4
2	berat, bmi, tinggi, umur, jenis kelamin	96.533
3	berat, bmi, tinggi, umur, gakin	96.4
4	berat, bmi, tinggi, jenis kelamin, gakin	95.733
5	berat, bmi, umur, jenis kelamin, gakin	96.266
6	berat, tinggi, umur, jenis kelamin, gakin	95.6
7	bmi, tinggi, umur, jenis kelamin, gakin	86.133
8	berat,bmi,tinggi,umur	96.666
9	berat,bmi,tinggi,jenis kelamin	95.866
10	berat,bmi,umur,jenis kelamin	96.266
11	berat,tinggi,umur,jenis kelamin	95.734
12	bmi,tinggi,umur,jenis kelamin	85.6
13	berat,bmi,tinggi	95.733
14	berat,bmi,umur	96.4
15	berat,tinggi,umur	95.866
16	bmi,tinggi,umur	86.133
17	berat,bmi	95.2
18	berat,umur	96
19	bmi,umur	81.866
20	berat	95.066
21	umur	75.6

Berdasarkan tabel 4.5, model yang dipilih sebagai model terbaik dari masing-masing bagian adalah model 1, 2, 8, 14, 17, dan 20. Dari keenam model tersebut dapat dilihat bagaimana pengaruh keberadaan parameter tidak independen (bmi, tinggi, jenis kelamin, umur) dan parameter independen (gakin). Pengaruh parameter tidak independen dan parameter independen ini dapat dilihat dari nilai akurasi masing-masing model.

1. Model 1 yang menggunakan semua parameter ketika dibandingkan dengan model 2 yang menghilangkan parameter jenis kelamin menunjukkan bahwa gakin memiliki pengaruh sebesar 0,13334% terhadap status gizi.
2. Model 8 dengan menghilangkan parameter jenis kelamin mengalami peningkatan 0,13334% dari model 2. Hal ini menunjukkan parameter jenis kelamin memiliki pengaruh 0,13334% terhadap status gizi.
3. Model 14 dengan menghilangkan parameter tinggi mengalami penurunan 0,26668% dari model 8. Hal ini menunjukkan bahwa parameter tinggi memiliki pengaruh 0,26668% terhadap status gizi.
4. Model 18 dengan menghilangkan parameter bmi mengalami penurunan akurasi 0,4% dari model 14. Hal

ini menunjukkan bahwa parameter bmi memiliki pengaruh 0,4% terhadap status gizi.

5. Model 20 dengan menghilangkan parameter umur mengalami peningkatan akurasi 0,93334% dari model 18. Hal ini menunjukkan bahwa parameter umur memiliki pengaruh 0,93334% terhadap status gizi.

Tabel 4.6 Model Data pada Skenario 80%:20%

Model	Parameter yang digunakan	Akurasi
1	berat, bmi, tinggi, umur, jenis kelamin, gakin	95,7
2	berat, bmi, tinggi, umur, jenis kelamin	95,8
3	berat, bmi, tinggi, umur, gakin	95,8
4	berat, bmi, tinggi, jenis kelamin, gakin	95,4
5	berat, bmi, umur, jenis kelamin, gakin	95,7
6	berat, tinggi, umur, jenis kelamin, gakin	94,9
7	bmi, tinggi, umur, jenis kelamin, gakin	87,3
8	berat, bmi, tinggi, umur	95,9
9	berat, bmi, tinggi, jenis kelamin	95,4
10	berat, bmi, umur, jenis kelamin	95,7
11	berat, tinggi, umur, jenis kelamin	95
12	bmi, tinggi, umur, jenis kelamin	86,4
13	berat, bmi, tinggi	95,5
14	berat, bmi, umur	95,9
15	berat, tinggi, umur	95,1
16	bmi, tinggi, umur	85,9
17	berat, tinggi	95,4
18	berat, umur	95,2
19	tinggi, umur	82.7
20	berat	94,5
21	tinggi	73,7

Berdasarkan tabel 4.6, model 1, 2, 8, 15, 17, dan 20 merupakan model yang dipilih sebagai model terbaik dari masing-masing bagian. Seperti halnya pada skenario 60%:40%, pada skenario 80%:20% ini juga terlihat bagaimana pengaruh parameter tidak independen dan parameter independen terhadap akurasi model. Pengaruh parameter tersebut dapat dijelaskan sebagai berikut.

1. Model 1 yang terdiri dari semua parameter jika dibandingkan dengan model 2 dengan menghilangkan parameter gakin mengalami peningkatan 0,1%. Hal ini menunjukkan bahwa gakin memiliki pengaruh 0,1% terhadap status gizi.
2. Model 8 dengan menghilangkan parameter jenis kelamin mengalami peningkatan akurasi dari model 2 sebesar 0,1%. Hal ini menunjukkan bahwa parameter jenis kelamin memiliki pengaruh 0,1% terhadap status gizi.
3. Model 14 dengan menghilangkan parameter tinggi tidak mengalami perubahan akurasi dari model 8. Hal ini menunjukkan parameter tinggi tidak memiliki pengaruh terhadap status gizi.
4. Model 18 dengan menghilangkan parameter bmi mengalami peningkatan 0,7% dari model 14. Hal ini menunjukkan parameter bmi memiliki pengaruh 0,7% terhadap status gizi.
5. Model 20 dengan menghilangkan parameter umur mengalami penurunan akurasi 0,7% dari model 18. Hal ini menunjukkan parameter umur memiliki pengaruh 0,7% terhadap status gizi.

4.3. Hasil dan Evaluasi

4.3.1. Penentuan Model Terbaik

Dalam menentukan model terbaik, dianalisa berdasarkan akurasi, waktu, dan akuisisi dari model. Model yang dipilih dilihat dari nilai akurasi, kesederhanaan model, waktu eksekusi, dan akuisisi data. Akuisisi dari model dilihat dari komponen parameter yang ada di dalam model tersebut. Akusisi parameter sendiri merupakan penentuan parameter lapangan yang cocok dengan daerah survey. Dalam kasus ini, parameter gakin kurang cocok digunakan, karena keadaan gakin pada balita sulit untuk ditentukan. Sedangkan untuk parameter berat, tinggi, umur, jenis kelamin, dan bmi, parameter tersebut saling terkait satu dengan yang lain, sehingga akuisisi data parameter tersebut tetap diperlukan sebagai data awal balita.

Tabel 4.7 Model Data pada Skenario 60%:40%

Model	Parameter	Training		Testing	
		Akurasi	Waktu	Akurasi	Waktu
1	berat,BMI,tinggi,umur,jenis kelamin, gakin	96,4	0,41	94	0,157
2	berat,BMI,tinggi,umur,jenis kelamin	96,53334	0,281	94	0,147
3	berat,BMI,tinggi,umur,gakin	96,4	0,282	94,2	0,147
4	berat,BMI,tinggi,jenis kelamin,gakin	95,73332	0,293	93,4	0,122
5	berat,BMI,umur,jenis kelamin,gakin	96,26666	0,323	94,2	0,141
6	berat,tinggi,umur,jenis kelamin,gakin	95,6	0,315	92	0,143
7	BMI,tinggi,umur,jenis kelamin,gakin	86,13334	0,319	76,6	0,138
8	berat,BMI,tinggi,umur	96,66668	0,276	94,4	0,128
9	berat,BMI,tinggi,jenis kelamin	95,86666	0,225	93,6	0,125
10	berat,BMI,umur,jenis kelamin	96,26666	0,247	94,2	0,137
11	berat,tinggi,umur,jenis kelamin	95,73334	0,283	92	0,138
12	BMI,tinggi,umur,jenis kelamin	85,6	0,29	76,8	0,129
13	berat,BMI,tinggi	95,73332	0,206	93,4	0,096
14	berat,BMI,umur	96,4	0,237	94,4	0,121
15	berat,tinggi,umur	95,86666	0,237	92	0,121
16	BMI,tinggi,umur	86,13332	0,25	76,6	0,116
17	berat,BMI	95,2	0,182	94,2	0,089
18	berat,umur	96	0,211	93	0,108
19	BMI,umur	81,86668	0,214	81	0,109
20	berat	95,06666	0,14	93,4	0,082
21	umur	76.6667	0.185	66	0.096

Berdasarkan tabel 4.7 di atas, pada data training 60%, model 8 memiliki akurasi paling tinggi, yaitu 96,66668%. Namun, model 14 memiliki nilai akurasi yang tidak signifikan dibandingkan model 8 yaitu dengan akurasi 96,4%. Dari segi waktu, model 14 juga memiliki waktu yang lebih singkat dibandingkan model 8. Akusisi data dari model 14 cocok untuk digunakan dalam klasifikasi status gizi. Pada data *training* dengan presentase 60% ini, model 14 merupakan model yang paling baik.

Pada data testing, model 8 dan model 14 memiliki nilai akurasi yang paling tinggi, yaitu 94,4%. Model 17 memiliki nilai akurasi tinggi yang tidak berbeda signifikan dengan model 8 dan 14 yaitu sebesar 94,2%. Namun, model 17 ketika dihitung menggunakan data training memiliki akurasi yang tidak maksimal. Dari segi akuisisi data, model 14 memiliki parameter yang cocok unuk klasifikasi status gizi. Meskipun dalam menentukan BMI dibutuhkan jenis kelamin dan tinggi, namun dalam penentuan status gizi hanya digunakan parameter berat,

BMI, dan umur saja. Oleh karena itu, model 14 dijadikan model terbaik untuk kasus skenario data 60%:40%.

Tabel 4.8 Model Data pada Skenario 80%:20%

Model	Parameter	Training		Testing	
		Akurasi	Waktu	Akurasi	Waktu
1	berat,bmi,tinggi,umur,jenis kelamin,gakin	95,7	0,473	94,8	0,051
2	berat,bmi,tinggi,umur,jenis kelamin	95,8	0,465	94,8	0,057
3	berat,bmi,tinggi,umur,gakin	95,8	0,546	94	0,053
4	berat,bmi,tinggi,jenis kelamin,gakin	95,4	0,445	94	0,048
5	berat,bmi,umur,jenis kelamin,gakin	95,7	0,54	93,6	0,055
6	berat,tinggi,umur,jenis kelamin,gakin	94,9	0,528	92,8	0,055
7	bmi,tinggi,umur,jenis kelamin,gakin	87,3	0,534	75,6	0,057
8	berat,bmi,tinggi,umur	95,9	0,432	94,4	0,056
9	berat,bmi,tinggi,jenis kelamin	95,4	0,383	94	0,056
10	berat,bmi,umur,jenis kelamin	95,7	0,434	93,6	0,058
11	berat,tinggi,umur,jenis kelamin	95	0,43	93,2	0,057
12	bmi,tinggi,umur,jenis kelamin	86,4	0,466	75,2	0,06
13	berat,bmi,tinggi	95,5	0,314	93,6	0,045
14	berat,bmi,umur	95,9	0,366	93,6	0,058
15	berat,tinggi,umur	95,1	0,377	94	0,06
16	bmi,tinggi,umur	85,9	0,363	76,4	0,052
17	berat,bmi	95,4	0,247	93,6	0,041
18	berat,umur	95,2	0,3	92,8	0,048
19	bmi,umur	82.7	0.298	81.2	0.054
20	berat	94,5	0,196	94	0,043
21	umur	73,7	0,204	68	0,046

Berdasarkan tabel 4.8, pada data training, model 8 dan 14 memiliki akurasi yang paling tinggi yaitu 95,9%. Jika dilihat waktu eksekusinya, model 14 memiliki waktu yang relatif lebih singkat daripada model 8. Dari akusisi data, model 14 terdiri dari parameter yang cocok untuk melakukan klasifikasi. Oleh karena itu, pada skenario data training 80% ini model yang terbaik ditunjukkan oleh model 14.

Pada data testing 20%, model yang memiliki akurasi paling tinggi merupakan model 1 dan model 2 dengan nilai akurasi sebesar 94,8%. Meskipun model 8 memiliki akurasi yang tidak signifikan, namun waktu eksekusi model 8 tidak berbeda jauh dengan model 2. Oleh karena itu, pada skenario data 80%:20% model yang paling baik ditunjukkan pada model 2 dengan parameter berat, BMI, tinggi, umur, dan jenis kelamin. Dari segi akuisisi data, model 2 cocok untuk pengklasifikasian status gizi.

4.3.2. Pengujian WEKA

Pengujian data menggunakan software WEKA hanya dilakukan pada pengujian dengan skenario data testing 40%. Dalam software WEKA juga telah dilakukan lima kali percobaan untuk setiap modelnya, sehingga dihasilkan perbandingan nilai akurasi pada sistem dan WEKA. Jika pada kedua sistem menunjukkan model terbaik yang sama, maka sistem dapat dikatakan berhasil.

Tabel 4.9 Perbandingan Hasil Akurasi *Testing*

MODEL	Hasil Akurasi Sistem (%)	Hasil Akurasi WEKA
1	94	94
2	94	94
3	94,2	94,4
4	93,4	93,6
5	94,2	94,4
6	92	92
7	76,6	76,6
8	94,4	94,4
9	93,6	94
10	94,2	94,2
11	92	92,4
12	76,8	76,4
13	93,4	93,8
14	94,4	94,4
15	92	92,2
16	76,6	77,2
17	94,2	94,2
18	93	90,4
19	81	81,8
20	93,4	93,4
21	66	68,8

Perbandingan hasil akurasi antara WEKA dan system yang dibuat dapat dilihat pada tabel 4.9. Jika dilihat dari perbandingan akurasinya, bisa dikatakan hasil akurasi system sama dengan hasil pada software WEKA. Terdapat perbedaan hasil akurasi pada beberapa model, namun perbedaan nilai yang dihasilkan tersebut tidak berbeda jauh. Untuk hasil model terbaik pada masing-masing bagian antara dua system tersebut ditunjukkan oleh model yang sama, yaitu model 1, 3, 8, 14, 17, dan 20. Model terbaik yang ditunjukkan oleh WEKA dan sistem yang dibuat sama, yaitu model 14 dengan akurasi 94,4%.

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa *Backward Feature Selection* dapat digunakan untuk pembentukan model berdasarkan nilai korelasi menggunakan metode *Cosine Similarity*. Berdasarkan akurasi, waktu eksekusi, dan akuisisi parameter yang digunakan, ditemukan model yang paling baik pada dua skenario percobaan. Pada skenario 60%:40%, model terbaik ditunjukkan oleh model yang terdiri dari parameter berat, bmi dan umur dengan nilai akurasi sebesar 94.4%. Sedangkan pada skenario 80%:20% model terbaik ditunjukkan pada model yang terdiri dari parameter berat, bmi, tinggi, umur, dan jenis kelamin dengan nilai akurasi 94,8%.

Penelitian ini menunjukkan bahwa korelasi parameter mempengaruhi hasil akurasi *Naïve Bayes Classifier*. Oleh karena itu, sebelum melakukan klasifikasi, perlu dilakukan perhitungan uji korelasi antar parameter terlebih dahulu. Selanjutnya dapat diketahui parameter apa yang bisa dihilangkan, sehingga mendapatkan model yang paling sederhana. Penelitian ini memunculkan asumsi baru, bahwa dalam *Naive Bayes Classifier*, sebelum melakukan uji korelasi antar parameter, akan lebih baik dilakukan uji korelasi parameter dengan atribut kelas hasil klasifikasi. Hal ini dikarenakan hubungan parameter dan kelas hasil klasifikasi mempengaruhi pembentukan model. Model terbaik pada skenario 80%:20% diperoleh dengan penghilangan parameter gakin yang memiliki korelasi paling rendah.

Seperti halnya pada penelitian yang dilakukan oleh Domingos dan Pazzani, penelitian ini juga menunjukkan performa yang baik meskipun terdiri dari parameter dependen. Bahkan, dalam penelitian ini model yang terbaik ditunjukkan oleh model yang terdiri dari parameter yang dependen.

Cosine Similarity tidak memiliki batasan untuk korelasinya, sehingga untuk penelitian selanjutnya penulis memberikan saran dalam uji korelasi *Spearman*. Fungsi *Spearman* tersebut memiliki batas untuk koefesien korelasi sehingga lebih jelas dalam menentukan korelasi antar parameter.

6. DAFTAR PUSTAKA

[1] Berson, A., and Smith S. J. (2001). Data Warehousing, Data Mining, & OLAP. New York, NY : McGraw-Hill.

[2] Domingos, P., and Pazzani, M. (1997). On the optimality of the Simple Bayesian *Classifier* under Zero-One Loss. Machine Learning, 29, 103–130

[3] Shadiq, M. A. (2009). Keoptimalan *Naïve Bayes* dalam Klasifikasi. Bandung: Universitas Pendidikan Indonesia.

[4] Supariasa, D. N., Bakri, B., & Fajar, I. (2002). *Penilaian Status Gizi.* Jakarta: Penerbit Buku Kedokteran EGC.

[5] Kusumadewi, S. (2009). Klasifikasi Status Gizi Menggunakan *Naive Bayes*ian Classification. *CommIT, Vol. 3 No. 1* , 6-11.

[6] Abe, S. (2005). Modified Backward Feature Selection. *European Symposium on Artificial Neural Networks* , 163-168.

[7] *Ladha, L., & Deepa, T. (2011).* Feature Selection Methods and Algorithms. *International Journal on Computer Science and Engineering (IJCSE) ,* Vol. 3 No.5, 1787-1797.

[8] Susanto, S., & Sensuse, D. I. (2008, Vol.1 No.2). Pengklasifikasian Artikel Berita Berbahasa Indonesia secara Otomatis Menggunakan Naive Bayes Classifier. *Jurnal Ilmu Komputer dan Informasi*

[9] O'Connor, B. T. (2012, Maret 13). *Cosine similarity, Pearson correlation, and OLS coefficients*. Retrieved Januari 7, 2015, from AI and Social Science - Brendan O'Connor: http://brenocon.com/blog/2012/03/cosine-similarity-pearson-correlation-and-ols-coefficients/

[10] Marmanis H, Babenko D. 2009. *Algorithms of the Intelligent Web*. Greenwich(UK): Manning Publ

[11] Nugroho, Bhuono Agung. 2005. *Strategi Jitu Memilih Metode Statistik Penelitian dengan SPSS*. Yogyakarta : Andi

[12] Witten, I. H., Frank, E., & Mark, A. H. (2011). *Data Mining : Practical Machine Learning Tools and Techniques 3rd Edition.* Burlington: Elsevier.

[13] Kementrian Kesehatan RI. (2011). *Standar Antropometri Penilaian Status Gizi Anak.* Jakarta: Menteri Kesehatan.