METODE ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK KLASIFIKASI TINGKAT KESEJAHTERAAN KESEHATAN MASYARAKAT PEKANBARU

KARYA ILMIAH

OLEH

MILA ANGGRAINI
NIM 1603137004

PROGRAM STUDI S1 STATISTIKA
JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS RIAU
2021

Repository University of Riau
https://repository.unri.ac.id
Hak cipta milik Universitas Riau
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository University of Riau
https://repository.unri.ac.id
© Hak cipta milik Universitas Riau

METODE ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK KLASIFIKASI TINGKAT KESEJAHTERAAN KESEHATAN MASYARAKAT PEKANBARU

Mila Anggraini, Arisman Adnan

Program Studi S1 Statistika
Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Riau
Kampus Bina Widya, Pekanbaru 28293

mila.anggraini7004@student.unri.ac.id

ABSTRACT

The level of public health welfare in a country can determine the quality and circumstances of the country. The purpose of this study is to classify the level of public health welfare of Pekanbaru City in 2019. The study used C4.5 and Naïve Bayes algorithms with k-fold cross validation to predict the accuracy of classification and performance evaluation measure as evaluation of both models. Performance evaluation measure results with k-fold cross validation show that models with Naïve Bayes have better classification results than C4.5 algorithm models. This is because Naïve Bayes' accuracy, precision, sensitivity and specificity are greater than the C4.5 algorithm by 100%. This is also because Naïve Bayes obtained the result from the probability value of each attribute being free of each other.

***Keyword**: C4.5 algorithm, Naïve Bayes, k-fold cross validation, performance evaluation measure, health wellbeing level.*

ABSTRAK

Tingkat kesejahteraan kesehatan masyarakat di suatu negara dapat menentukan kualitas dan keadaan dari negara tersebut. Tujuan penelitian ini untuk mengklasifikasikan tingkat kesejahteraan kesehatan masyarakat Kota Pekanbaru tahun 2019. Penelitian ini menggunakan algoritma C4.5 dan *Naïve Bayes* dengan *k-fold cross validation* untuk memprediksi ketepatan klasifikasi serta *performance evaluation measure* sebagai evaluasi dari kedua model. Hasil *performance evaluation measure* dengan *k-fold cross validation* menunjukkan bahwa model dengan *Naïve Bayes* memiliki hasil klasifikasi yang lebih baik daripada model algoritma C4.5. Hal ini dikarenakan nilai akurasi, presisi, sensitivitas dan spesifisitas *Naïve Bayes* lebih besar dari algoritma C4.5 sebesar 100%. Hal ini juga dikarenakan *Naïve Bayes* perolehan hasilnya dari nilai probabilitas dari setiap atribut yang bebas antara satu sama lain.

Kata kunci: Algoritma C4.5, *Naïve Bayes*, *k-fold cross validation*, *performance evaluation measure*, tingkat kesejahteraan kesehatan.

Repository Universitas Riau
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.

1. PENDAHULUAN

Terwujudnya kesejahteraan masyarakat merupakan cita-cita dan tujuan utama bagi suatu negara. Tingkat kesejahteraan masyarakat di suatu negara dapat menentukan kualitas dan keadaan negara tersebut. Pemerintah Provinsi Riau terus berusaha untuk meningkatkan penyediaan fasilitas kesehatan dan sumber daya manusia di bidang kesehatan. Salah satu upaya yang dilakukan Pemerintah Provinsi Riau tepatnya di Kota Pekanbaru dengan melakukan berbagai program dam kegiatan yang ditujukan untuk meningkatkan derajat kesehatan masyarakat (Dinas Kesehatan, 2018).

Tingkat kesejahteraan kesehatan masyarakat dapat digambarkan dengan jumlah sarana dan prasarana kesehatan, keluhan kesehatan, angka kesakitan, angka penolong kelahiran terakhir dan rata-rata lama diberi ASI. Gambaran ini dapat dianalisis dengan metode data *mining* yang memprediksi suatu kelas dari sekumpulan data dengan mengenali pola dari data baik data numerik maupun kategori dalam bentuk klasifikasi.

Pengklasifikasian dalam data *mining* dapat dilakukan dengan metode Algoritma C4.5 maupun *Naïve Bayes*. Algoritma C4.5 dapat menangani data numerik dan diskrit yang mampu melakukan klasifikasi sekaligus menunjukkan hubungan antar atribut (Aprilia et al. 2013). *Naïve Bayes* memprediksi probabilitas keanggotaan suatu kelas dengan mengasumsikan bahwa pengaruh nilai atribut pada kelas yang diberikan tidak tergantung pada nilai atribut lainnya atau berdasar pada teorema *Bayes* (Han et al. 2012).

Penelitian sebelumnya mengenai identifikasi tumbuh kembang anak balita di Puskesmas Kartasura menggunakan perbandingan algoritma C4.5 dan *Naïve Bayes* (Listiana et al. 2015). Penelitian ini mengungkapkan bahwa metode *Naive Bayes* lebih baik digunakan dari pada metode algoritma C4.5 dengan nilai total 250.67% untuk algoritma C4.5 dan 258.03% untuk *Naive Bayes*. Penelitian lainnya mengenai klasifikasi data kesehatan ibu di Kabupaten Jammu dan Kashmir, India pada tahun 2016-2017 menggunakan metode *Naïve Bayes* (Kour et al. 2019). Pengklasifikasian kabupaten menjadi kabupaten prioritas dan non prioritas dapat dikelompokkan dengan metode berdasarkan data kesehatan ibu dan kinerja metode ini menunjukkan hasil yang baik.

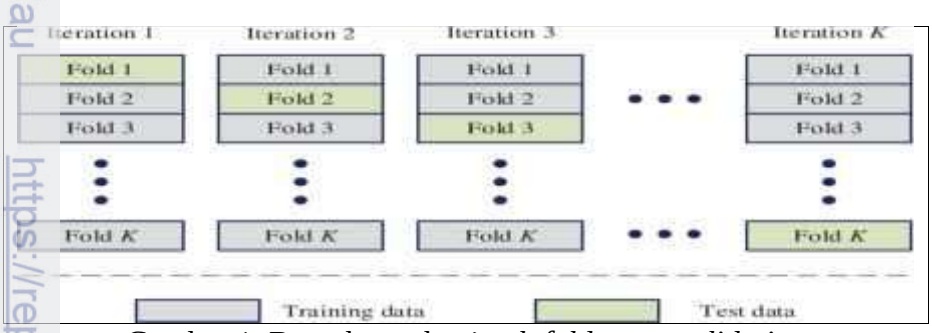
2. ALGORITMA C4.5 DAN NAÏVE BAYES

Data *mining* merupakan suatu proses pemeriksaan sekumpulan besar data yang tersimpan dalam penyimpanan atau *database* dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika (Larose & Larose, 2014). Teknik dalam data mining memungkinkan untuk membuat sebuah model untuk mengetahui pola data yang lain selain dalam basis data yang tersimpan. Ada beberapa teknik yang terdapat dalam data *mining* (Pramudiono, 2013), yaitu:

a. *Association rule mining*, proses dari asosiasi yang muncul pada setiap kejadian.
b. *Clustering*, pemberian label pada kelas data yang belum diketahui dengan membaginya menjadi kelompok-kelompok dalam data *mining*.
c. Klasifikasi, teknik untuk menggambarkan kelas atau konsep dari suatu data.
d. Estimasi, menduga sebuah nilai yang tidak diketahui.
e. Prediksi, memperkirakan atau *forecasting* suatu kejadian yang belum terjadi.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository Universitas Riau https://repository.unri.ac.id
© Hak cipta milik Universitas Riau

Model pengklasifikasian dapat memprediksi label kelas kategorikal (diskrit, tidak berurutan). Data *mining* membuat model klasifikasi dengan memeriksa data yang sudah diklasifikasikan dan secara induktif menemukan pola prediksi. Dalam klasifikasi terdapat dua proses yaitu *training* dan *testing*. Pada proses *training* sebagian datanya yang diketahui kelasnya dilatih dan menghasilkan sebuah model prediksi. Pada proses *testing*, dari model prediksi sebelumnya diuji dengan sebagian data lainnya guna mengetahui keakuratan dari model prediksi tersebut. Pembagian data *training* dan data *testing* ditentukan dari teknik *k-fold cross validation* yang merupakan prosedur *resampling* untuk mengevaluasi model klasifikasi pada data yang memiliki sampel terbatas (James et al. 2013), dengan bentuk yang disajikan pada Gambar 1.



Gambar 1. Bentuk pembagian *k-fold cross validation*

Algoritma C4.5 merupakan bagian dari algoritma *decision tree* yang dikembangkan oleh Quinlan pada tahun 1996-an (Witten et al. 2017). *Decision tree* merupakan model prediksi menggunakan struktur berhierarki dengan bentuk pohon yang paling dasar yang disebut ID3 atau Iterative Dichotomiser 3. Beberapa tahapan dalam membuat *decision tree* dengan algoritma C4.5 (Kane et al. 2016), adalah:

a. Menentukan atribut sebagai akar dari pohon dengan memilih nilai *gain ratio* tertinggi yang akan menjadi akar pertama (Khun &Johnson, 2013). Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai *entropy*, *gain* dan *split info*. *Entropy* adalah parameter yang digunakan untuk mengukur tingkat keberagaman (heterogenitas) dari sekumpulan data. Bentuk persamaan untuk mendapatkan nilai *entropy* disajikan sebagai berikut:

$$Entropy\,(S) = \sum_{i=1}^{n} (-\,p_i)\log_2 p_i,$$

dengan S menyatakan kumpulan data, n jumlah nilai pada atribut target (jumlah kelas) dan p_i proporsi atau rasio antara jumlah sampel di kelas *i* dengan jumlah semua sampel pada S.

Gain merupakan pembobotan pada suatu data dengan menggunakan maksimal *entropy* yang disajikan pada persamaan sebagai berikut:

$$Gain\,(S,A) = Entropy\,(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy\,(S_i),$$

dengan S menyatakan kumpulan data, A atribut, n banyaknya partisi atribut *A*, $|S_i|$ banyaknya data pada partisi ke $-i$ dan $|S|$ banyaknya data dalam S.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository Universitas Riau
http://repository.unri.ac.id
© Hak cipta milik Universitas Riau

Gain dimodifikasi menjadi *gain ratio* untuk mengurangi bias atribut yang memiliki banyak cabang yang bernilai besar jika data menyebar rata dan bernilai kecil jika semua data masuk ke dalam satu cabang (Mohammed et al. 2017), dengan persamaan seperti yang disajikan pada persamaan (1).

$$Gain\ Ratio\ (S,A) = \frac{Gain\ (S,A)}{Split\ Info\ (S,A)}, \tag{1}$$

dengan *split info* menyatakan *entropy* atau informasi potensial yang dihitung seperti persamaan sebagai berikut:

$$Split\ Info\ (S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S}.$$

b. Ulangi langkah pertama untuk percabangan dari akar pertama hingga semua *record* terpartisi.
c. Proses partisi akan berhenti ketika semua atribut pada cabang memiliki kelas yang sesuai, tidak ada atribut yang bisa dibagi lagi dan tidak ada data *record* pada cabang yang kosong.

Setelah *decision tree* terbentuk, terdapat aturan (*rule set*) yang membentuk kondisi berupa *if-then* (Marlina et al. 2016), dengan bentuk seperti yang disajikan pada Tabel 1.

Tabel 1. *Rule set decision tree*

A	*B*	*A* dan *B*
False	False	False
False	True	False
True	False	False
True	True	True

'*A*', '*B*' merupakan nilai atribut pada pohon keputusan, '*A* dan *B*' merupakan pemberian label kelas yang sesuai.

Naïve Bayes adalah metode klasifikasi yang berdasarkan teori probabilitas dan teorema *Bayes* dengan asumsi bahwa setiap variabel bebas antara satu sama lain atau independen (Pradeep & Naveen, 2018). Dalam data *mining*, asumsi dari teorema *Bayes* untuk metode *Naïve Bayes* dengan algoritma sebagai berikut:

a. Variabel *D* merupakan kumpulan data memiliki label kelas. Setiap data berdimensi *n* yang dinyatakan sebagai $X = (x_1, x_2, x_3, \ldots, x_n)$ yang diperoleh dari *n* atribut $A_1, A_2, A_3, \ldots, A_n$.
b. Misalkan terdapat *m* kelas $(C_1, C_2, C_3, \ldots, C_n)$ pada data *X*, *Naïve Bayes* memprediksi bahwa data *X* berada dalam kelas C_i jika dan hanya jika

$$P(C_i \mid X) > P(C_j \mid X) \text{ untuk } i \le j \le m,\ j \ne i,$$

dengan kata lain kelas C_i pada *Naïve Bayes* memaksimalkan nilai $P(C_i \mid X)$ yang disebut sebagai *maximum posteriori hypothesis*. *Naïve Bayes* memiliki asumsi penyederhanaan atau *naïve* (Han et al. 2012), sehingga *Naïve Bayes* berbentuk seperti yang disajikan pada persamaan sebagai berikut:

$$P(C_i \mid X) = \frac{P(X \mid C_i) P(C_i)}{P(X)},$$

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository University of Riau
https://repository.unri.ac.id
© Hak cipta milik Universitas Riau

dengan X menyatakan data dengan kelas yang tidak diketahui, C_i hipotesis data dari kelas X, $P(C_i \mid X)$ hipotesis C_i berdasarkan kondisi X, $P(C_i)$ hipotesis C_i, $P(X \mid C_i)$ hipotesis X berdasarkan kondisi C_i dan $P(X)$ hipotesis X. Nilai *P(X)* bersifat konstan untuk setiap kelas pada C_i, sehingga $P(X \mid C_i)P(C_i)$ dimaksimalkan seperti yang disajikan pada persamaan (2).

$$\begin{aligned} P(C_i \mid X) &= P(C_i)\prod_{k=1}^{n} P(x_k \mid C_i), \\ &= P(x_1 \mid C_i)P(x_{k2} \mid C_i)P(x_3 \mid C_i)\ldots P(x_n \mid C_i)P(C_i). \end{aligned} \tag{2}$$

c. Nilai $P(C_i)$ pada persamaan (2) dihitung dengan perhitungan jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut, dengan perhitungan seperti yang disajikan pada persamaan sebagai berikut:

$$P(C_i) = \frac{n(C_i)}{N},$$

dengan $P(C_i)$ menyatakan probabilitas kejadian kelas C_i, $n(C_i)$ jumlah kelas C_i yang dikategori dan N jumlah keseluruhan kelas C_i.

d. Nilai probabilitas $P(X \mid C_i)$ yakni probabilitas bersyarat tiap-tiap variabel terhadap kelas, dihitung seperti yang disajikan pada persamaan sebagai berikut:

$$P(X \mid C_i) = \frac{P(C_i \cap X)}{P(C_i)}.$$

Apabila terdapat probabilitas yang bernilai nol, maka digunakan teknik perhitungan *Laplace Correction* dengan menambah nilai satu pada setiap data yang dihitung (Mohammed et al. 2017), dengan perhitungan seperti yang disajikan pada persamaan sebagai berikut:

$$p_i = \frac{m_i + 1}{n + k}, \tag{3}$$

dengan p_i menyatakan probabilitas dari atribut m_i, m_i jumlah data dalam kelas dari atribut m_i, n jumlah data dan k jumlah kelas dari atribut m_i.

e. Prediksi untuk label kelas dari data X dengan menghitung nilai probabilitas *posterior* dengan persamaan (2) pada setiap kelas terhadap atribut dan nilai data tertinggi diberi label kelas.

Performance Evaluation Measure (*PEM*) merupakan suatu pengukuran performa dari model pada kasus yang menghasilkan evaluasi dari model yang digunakan (Bramer, 2013). *PEM* digambarkan dalam *confusion matrix* yang berisi hasil pengujian dari suatu model atau sistem berbentuk tabel diagonal yang disajikan pada Tabel 2.

Tabel 2. *Confusion matrix*

		Predicted Values		
		A	*B*	*C*
Actual Values	*A*	*AA*	*AB*	*AC*
	B	*BA*	*BB*	*BC*
	C	*CA*	*CB*	*CC*

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository University of Riau https://repository.unri.ac.id/ © Hak cipta milik Universitas Riau

Dari 9 istilah dalam *confusion matrix* dapat diperoleh nilai akurasi, presisi, sensitivitas dan spesifisitas. Nilai akurasi menggambarkan seberapa akurat sistem atau model dapat mengklasifikasikan data secara benar (Han et al. 2012), dengan perhitungan seperti yang disajikan pada persamaan (4).

$$Ak = \frac{AA + BB + CC}{AA + AB + AC + BA + BB + BC + CA + CB + CC}. \quad (4)$$

Nilai presisi mencerminkan akurasi klasifikasi yang mengukur kinerja dari klasifikasi model (Li et al. 2015), dengan perhitungan seperti yang disajikan pada persamaan (5).

$$Pres = \frac{\frac{AA}{AA + BA + CA} + \frac{BB}{AB + BB + CB} + \frac{CC}{AC + BC + CC}}{\text{Total Kelas}}. \quad (5)$$

Sementara itu, sensitivitas menunjukkan derajat keandalan model dalam mendeteksi data berlabel positif dengan benar oleh sistem atau model (Shan et al. 2010), dengan perhitungan seperti yang disajikan pada persamaan (6).

$$Sens = \frac{\frac{AA}{AA + AB + AC} + \frac{BB}{BA + BB + BC} + \frac{CC}{CA + CB + CC}}{\text{Total Kelas}}. \quad (6)$$

Performa selanjutnya mengenai spesifisitas yang menunjukkan derajat keandalan model dalam memprediksi data berlabel negatif dengan benar (Han et al. 2012), dengan perhitungan seperti yang disajikan pada persamaan (7).

$$Spes = \frac{\frac{BB + CC}{BA + CA + BB + CC} + \frac{AA + CC}{AB + CB + AA + CC} + \frac{AA + BB}{AC + BC + AA + BB}}{\text{Total Kelas}}. \quad (7)$$

3. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini berupa data sekunder yang diperoleh dari Badan Pusat Statistik Provinsi Riau yang merupakan hasil Survei Sosial Ekonomi Nasional (Susenas) tahun 2019. Banyaknya data yang digunakan dalam penelitian ini sebanyak 215 data dan 12 variabel. Variabel yang digunakan, yakni variabel kesejahteraan kesehatan, keluhan kesehatan, jaminan kesehatan, merokok, imunisasi, persalinan, sumber air bersih, jenis jamban, ASI eksklusif, KB, berobat jalan dan rawat inap. Proses klasifikasi dalam penelitian ini menggunakan *software* R Studio versi 4.0.3. Langkah analisis:

a. Mempartisi data dengan *k-fold cross validation* sebanyak 10-*fold*.
b. Pemodelan algoritma C4.5 dengan 10-*fold* untuk menentukan *root* dari *decision tree* dengan nilai *gain ratio* tertinggi menjadi *root* pertama. *Root* yang terpilih dari atribut yang dihitung dari *gain ratio* ditentukan terlebih dahulu nilai *entropy* untuk mengukur keberagaman dari sekumpulan data yang di *training*.
c. Pemodelan pada *Naïve Bayes* dengan 10-*fold* untuk menentukan nilai maksimal dari probabilitas *posterior* pada kelas data. Penentuan dimulai dari menghitung probabilitas dari setiap kelas dan menghitung probabilitas bersyarat tiap-tiap variabel tehadap kelas sehingga terbentuk model terbaik dari *Naïve Bayes*.
d. Setelah model terbaik algoritma C4.5 dan *Naïve Bayes* diperoleh, kemudian dilakukan evaluasi prediksi klasifikasi melalui *performance evaluation measure*

dalam bentuk *confusion matrix*. *Performance evaluation measure* menghasilkan nilai akurasi, presisi, *recall* atau sensitivitas dan spesifisitas sebagai perbandingan dari kedua model.

4. TINGKAT KESEJAHTERAAN KESEHATAN MASYARAKAT KOTA PEKANBARU TAHUN 2019

Pengklasifikasian terhadap data tingkat kesejahteraan kesehatan masyarakat Pekanbaru menggunakan metode algoritma C4.5 dan *Naïve Bayes* dengan perbandingan melalui hasil *performance evaluation measure*. Pembagian data *training* dan data *testing* dari *k-fold cross validation* dengan *k* sebanyak 10-*fold*, membagi data tingkat kesejahteraan kesehatan masyarakat Pekanbaru menjadi 10 kali dan perulangan sebanyak 10 kali.

Dari 10-*fold cross validation* membentuk 10 jenis keputusan (*rule set*) berdasarkan hasil persamaan (1) untuk data *training* dari algoritma C4.5 yang disajikan pada Tabel 3.

Tabel 3. Hasil prediksi klasifikasi algoritma C4.5

K-fold	Benar	Salah	Jumlah	Akurasi
1	187	7	194	0.9639
2	186	7	193	0.9637
3	187	7	194	0.9639
4	188	6	194	0.9691
5	182	11	193	0.9430
6	187	7	194	0.9639
7	187	6	193	0.9689
8	189	4	193	0.9793
9	188	5	193	0.9741
10	186	6	192	0.9688

Dari Tabel 3 dapat dilihat bahwa kesalahan prediksi dari model algoritma C4.5 dengan 10-*fold cross validation* rata-rata 6 hingga 7 data dan terbesar 11 data, diperoleh tingkat akurasi rata-rata sebesar 96.59%.

Data *testing* dari algoritma C4.5 yang diperoleh berdasarkan hasil *rule set* data *training* di evaluasi menggunakan metrik evaluasi dari *confusion matrix* dalam bentuk *performance evaluation measure*. Jumlah klasifikasi algoritma C4.5 dengan data *testing* disajikan pada Tabel 4.

Tabel 4. Jumlah prediksi data *testing* algoritma C4.5

K-fold	Rendah	Sedang	Tinggi	Jumlah
1	3	9	9	21
2	4	10	8	22
3	5	7	9	21
4	3	12	6	21
5	3	7	12	22
6	3	12	6	21
7	2	12	8	22
8	2	10	10	22
9	3	4	8	22
10	3	9	9	21

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository Universitas Riau https://repository.unri.ac.id
© Hak cipta milik Universitas Riau

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository Universitas Riau
https://repository.unri.ac.id
© Hak cipta milik Universitas Riau

Berdasarkan data *testing* untuk setiap *k-fold cross validation* sebanyak 21 dan 22 data tingkat kesejahteraan kesehatan masyarakat Pekanbaru, didapatkan bahwa masyarakat dengan tingkat kesejahteraan kesehatan kelas rendah berkisar antara 2 hingga 5 data, kelas sedang berkisar antara 4 hingga 12 data dan kelas tinggi berkisar antara 6 hingga 12 data.

Berdasarkan persamaan (4), persamaan (5), persamaan (6) dan persamaan (7) untuk tiap-tiap representasi *performance evaluation measure* dari jumlah prediksi data *testing* algoritma C4.5 disajikan pada Tabel 5.

Tabel 5. Evaluasi data *testing* algoritma C4.5

K-fold	Akurasi	Presisi	Sensitivitas	Spesifisitas
1	0.9524	0.9630	0.9667	0.9744
2	0.8636	0.8417	0.8977	0.9244
3	0.8571	0.8296	0.9000	0.9305
4	0.9048	0.9444	0.9167	0.9394
5	0.8182	0.8889	0.8788	0.9048
6	0.9048	0.9444	0.9167	0.9394
7	0.8636	0.9028	0.8169	0.9138
8	0.8182	0.8667	0.7852	0.8889
9	0.9091	0.9280	0.9280	0.9459
10	0.9524	0.9630	0.9667	0.9744

Dari Tabel 5 dapat dilihat bahwa dengan menggunakan 10-*fold cross validation*, nilai terbesar terletak pada 1-*fold* dan 10-*fold*, terendah terletak pada 8-*fold*. Presentase nilai akurasi, presisi, sensitivitas dan spesifisitas pada 1-*fold* dan 10-*fold* sama, yakni 95.24%, 96.30%, 96.67% dan 97.44%. hal ini menunjukkan bahwa metode algoritma C4.5 sangat baik dalam mengklasifikasikan data tingkat kesejahteraan kesehatan masyarakat Pekanbaru tahun 2019.

Berdasarkan persamaan (2) dan persamaan (3) dari 10-*fold cross validation* membentuk 10 jenis prediksi probabilitas *posterior* untuk data *training* dari *Naïve Bayes* yang disajikan pada Tabel 6.

Tabel 6. Hasil prediksi klasifikasi *Naïve Bayes*

K-fold	Benar	Salah	Jumlah	Akurasi
1	191	3	194	0.9845
2	189	4	193	0.9793
3	188	6	194	0.9691
4	183	11	194	0.9433
5	185	8	193	0.9585
6	189	5	194	0.9742
7	188	5	193	0.9741
8	188	5	193	0.9741
9	189	4	193	0.9793
10	189	5	194	0.9742

Dari Tabel 6 dapat dilihat bahwa kesalahan prediksi model *Naïve Bayes* rata-rata 5 data dan terbesar 11 data dengan tingkat akurasi rata-rata sebesar 97.11% dari penggunaan 10-*fold cross validation*.

Klasifikasi *Naïve Bayes* dengan data *testing* digunakan untuk mengevaluasi model dengan *performance evaluation measure*. Jumlah klasifikasi *Naïve Bayes* dengan data *testing* berdasarkan hasil persamaan (2) dan persamaan (3) disajikan pada Tabel 7.

Tabel 7. Jumlah prediksi data *testing Naïve Bayes*

K-fold	Rendah	Sedang	Tinggi	Jumlah
1	3	9	9	21
2	4	10	8	22
3	3	10	8	21
4	3	11	7	21
5	3	13	6	22
6	3	12	6	21
7	3	13	6	22
8	3	10	9	22
9	3	10	9	22
10	4	9	8	21

Berdasarkan data *testing* untuk setiap *k-fold cross validation* sebanyak 21 dan 22 data tingkat kesejahteraan kesehatan masyarakat Pekanbaru, didapatkan bahwa masyarakat dengan tingkat kesejahteraan kesehatan kelas rendah berkisar antara 3 hingga 4 data, kelas sedang berkisar antara 9 hingga 13 data dan kelas tinggi berkisar antara 6 hingga 9 data.

Berdasarkan persamaan (4), persamaan (5), persamaan (6) dan persamaan (7) untuk tiap-tiap representasi *performance evaluation measure* dari jumlah prediksi data *testing Naïve Bayes* disajikan pada Tabel 8.

Tabel 8. Evaluasi data *testing Naïve Bayes*

K-fold	Akurasi	Presisi	Sensitivitas	Spesifisitas
1	0.9524	0.9630	0.9667	0.9744
2	0.9545	0.9167	0.9697	0.9825
3	1.0000	1.0000	1.0000	1.0000
4	0.9524	0.9697	0.9583	0.9697
5	0.9091	0.9487	0.9167	0.9394
6	0.9048	0.9444	0.9167	0.9394
7	0.8182	0.8675	0.8447	0.8853
8	1.0000	1.0000	1.0000	1.0000
9	0.9545	0.9630	0.9697	0.9762
10	0.9524	0.9167	0.9667	0.9815

Dari Tabel 8 dapat dilihat bahwa dengan menggunakan 10-*fold cross validation*, nilai terbesar terletak pada 3-*fold* dan 8-*fold*, terendah terletak pada 7-*fold*. Presentase pada 3-*fold* dan 8-*fold* sama-sama 100% untuk setiap nilai akurasi, presisi, sensitivitas dan spesifisitas. Hal ini menunjukkan bahwa metode *Naïve Bayes* sangat baik dan cocok dalam mengklasifikasikan data tingkat kesejahteraan kesehatan masyarakat Pekanbaru tahun 2019. Hal ini juga dikarenakan *Naïve Bayes* perolehan hasilnya dari nilai probabilitas dari setiap atribut yang bebas antara satu sama lain.

Berdasarkan pengklasifikasian algoritma C4.5 dan *Naïve Bayes* didapatkan hasil klasifikasi pada data tingkat kesejahteraan kesehatan masyarakat di kota Pekanbaru

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository University of Riau
https://repository.unri.ac.id
© Hak Cipta Dilindungi Universitas Riau

tahun 2019. Berikut diberikan perbandingan hasil klasifkasi dari algoritma C4.5 dan *Naïve Bayes* yang disajikan pada Tabel 9.

Tabel 9. Perbandingan klasifikasi algoritma C4.5 dan *Naïve Bayes*

Evaluasi Model	Algoritma C4.5	*Naive Bayes*
Akurasi	0.9524	1.0000
Presisi	0.9630	1.0000
Sensitivitas	0.9667	1.0000
Spesifisitas	0.9744	1.0000

Dari Tabel 9 dapat dilihat bahwa model *Naïve Bayes* memiliki hasil klasifikasi yang lebih baik daripada model algoritma C4.5. Hal ini dikarenakan baik nilai akurasi, presisi, sensitivitas maupun nilai spesifisitasnya lebih besar daripada algoritma C4.5. Hal ini juga dikarenakan *Naïve Bayes* perolehan hasilnya dari nilai probabilitas dari setiap atribut yang bebas antara satu sama lain, sedangkan algoritma C4.5 berdasarkan pada atribut yang memiliki nilai *gain ratio* tertinggi.

5. KESIMPULAN

Pengklasifikasian data tingkat kesejahteraan kesehatan masyarakat Pekanbaru tahun 2019 menggunakan algoritma C4.5 dan *Naïve Bayes* menunjukkan hasil klasifikasi kedua metode sangat baik. Hasil *performance evaluation measure* algoritma C4.5 dengan 10-*fold cross validation*, diperoleh nilai akurasi terbaik sebesar 95.24%, presisi terbaik sebesar 96.30%. sensitivitas terbaik sebesar 96.67 dan spesifisitas sebesar 97.44%. Hasil *performance evaluation measure Naïve Bayes* dengan 10-*fold cross validation*, diperoleh nilai terbaik sebesar 100% untuk setiap nilai akurasi, presisi, sensitivitas dan spesifisitas. Naïve Bayes lebih unggul dari algoritma C4.5 dalam klasifikasi data tingkat kesejahteraan kesehatan masyarakat Pekanbaru tahun 2019 karena hasil *performance evaluation measure* yang lebih besar dan perolehan hasilnya dari nilai probabilitas dari setiap atribut yang bebas antara satu sama lain.

DAFTAR PUSTAKA

Aprilia, D., Aji Baskoro, D., Ambarwati, L., & Wicaksana, I. W. S. (2013). *Belajar data mining dengan rapid minner*. Jakarta: Remi Sanjaya.

Bramer, M. (2013). *Principles of data mining*. London: Springer.

Dinas Kesehatan. (2018). *Profil kesehatan Kota Pekanbaru*. Pekanbaru: Dinas Kesehatan Kota Pekanbaru.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Data mining concepts and techniques*. New York: Morgan Kaufmann Publisher.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository University of Riau https://repository.unri.ac.id
© Hak Cipta milik Universitas Riau

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan Universitas Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin Universitas Riau.
Repository University of Riau https://repository.unri.ac.id
© Hak cipta milik Universitas Riau

Kane, S. N., Mishra, A., & Dutta, A. K. (2016). Analysis of data mining classification by comparison of C4.5 and ID3 algorithms. *Journal of Physics: Conference Series*, *755* (1), 1–9.

Kour, P., Shastri, S., Bhadwal, A. S., Kumar, S., Singh, K., Kumari, M., & Mansotra, V. (2019). Classification of maternal healthcare data using naive bayes. *International Journal of Computer Sciences and Engineering*, *7* (3), 388–394.

Khun, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining* (2nd Ed). New Jersey: John Wiley & Sons.

Li, L., Wu, Y., & Ye, M. (2015). Experimental comparisons of multi-class classifiers. *Informatica*, *39* (1), 71–85.

Listiana, M., Sudjalwo, & Gunawan, D. (2015). Comparison of decision tree algorithm (c4.5) and I bayes on data mining for identification growth and development child toddler. *Journal of Informatics Engineering and Computer Science*, *12* (3), 210–227.

Marlina, L., Lim, M., & Siahaan, U. A. P. (2016). Data mining classification comparison (Naïve Bayes and C4.5 algorithms). *International Journal of Engineering Trends and Technology*, *38* (7), 380–383.

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2017). *Machine learning algorithms and applications*. Boca Raton: CRC Press.

Pradeep, K. R., & Naveen, N. C. (2018). Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4.5 and I Bayes algorithms for healthcare analytics. *Procedia Computer Science*, *132* (1), 412–420.

Pramudiono. (2013). *Algoritma pembelajaran data mining*. Yogyakarta: Andi Offset.

Shan, Y., Milios, E. E., Roger, A. J., Blouin, C., & Susko, E. (2010). Automatic recognition of regions of intrinsically poor multiple alignment using machine learning. *Journal Bioinformatics and Computational Biology*, *15* (28), 482–483.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4nd Ed.). Cambridge: Morgan Kaufmann Publisher.