Perbandingan Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Penyakit Anak

Syamsul Bahri[1], Dwi Marisa Midyanti[2], Rahmi Hidayati[3]
Jurusan Sistem Komputer, Fakultas MIPA
Universitas Tanjungpura
Pontianak
[1]syamsul.bahri@siskom.untan.ac.id, [2]dwi.marisa@siskom.untan.ac.id, [3]rahmihidayati@siskom.untan.ac.id

***Abstrak*—Anak sangat rentan terhadap kuman penyakit, para orang tua pada umumnya sering tidak mengetahui gejala-gejala penyakit yang muncul pada tubuh anak. Informasi mengenai gejala-gejala penyakit pada tubuh anak sangat penting untuk diketahui orang tua agar orang tua mampu melakukan tindakan yang tepat pada awal gejala. Salah satu langkah awal yang dapat dilakukan dalam mendiagnosa penyakit pada anak adalah dengan melakukan klasifikasi berdasarkan gejala-gejala yang sering timbul. Terdapat beberapa algoritma klasifikasi, diantaranya yang sering digunakan yaitu Naive Bayes dan Decision Tree C4.5. Penelitian ini melakukan perbandingan kedua algoritma tersebut untuk klasifikasi penyakit anak. Dari hasil perbandingan menunjukkan bahwa algoritma terbaik dengan akurasi sebesar 90.00% yaitu Decision Tree C4.5. Sementara algoritma Naive Bayes memperoleh tingkat akurasi sebesar 89.58%.**

Kata kunci—Penyakit Anak, Decision Tree, C4.5, Naive Bayes, Perbandingan, Klasifikasi

I. PENDAHULUAN

Kesehatan merupakan bagian kehidupan manusia yang sangat penting, sebab setiap manusia dapat terganggu kesehatannya. Tidak terkecuali anak yang mudah terkena penyakit, sehingga sangat penting bagi orang tua untuk mendapatkan segala bentuk informasi tentang penyakit anak yang sedang diderita.

Kebanyakan orang tua pada umumnya sering tidak mengetahui penyakit yang gejala-gejalanya timbul pada tubuh anak. Informasi mengenai gejala-gejala penyakit pada tubuh anak sangat penting untuk diketahui orang tua agar orang tua mampu melakukan tindakan awal yang tepat. Hal ini dikarenakan jika terjadi kesalahan ataupun keterlambatan dalam mengenali gejala-gejala serta jenis penyakit pada anak, dapat menyebabkan kesalahan ataupun keterlambatan dalam penanganan pengobatannya hingga dapat menyebabkan kematian.

Penyakit yang sering terjadi pada anak, diantaranya seperti penyakit infeksi saluran napas akut (ISPA), diare, demam berdarah dengue (DBD), demam tifoid (tipes), Pneumonia, varisella (cacar air), dan masih banyak lagi. Salah satu masalah kesehatan yang sering terjadi di masyarakat pada negara berkembang seperti di Indonesia yaitu penyakit diare yang umumnya juga sering diderita oleh anak-anak. Diare merupakan penyebab kematian terbanyak kedua setelah Pneumonia. Departemen Kesehatan melakukan survei morbiditas yang menunjukkan hasil terdapat peningkatan jumlah penderita dari tahun 2009 sebanyak 53.854 hingga tahun 2010 mencapai 54.612 [1].

Kajian mengenai bagaimana mengenali gejala-gejala penyakit pada anak sangat penting untuk dilakukan, mengingat tingginya tingkat kasus penyakit pada anak dan minimnya pemahaman orang tua mengenai penyakit pada anaknya. Penerapan teknik analisa konvensional secara manual tidak efektif digunakan untuk mendiagnosa penyakit. Seiring dengan perkembangan teknologi dan sistem informasi, proses analisa dalam diagnosa penyakit menggunakan sistem berbasis komputer dengan pengetahuan medis menjadi semakin penting [2].

Salah satu cara yang bisa dilakukan dalam mendiagnosa penyakit pada anak adalah dengan melakukan klasifikasi berdasarkan gejala-gejala yang sering timbul. Klasifikasi merupakan proses menemukan fungsi atau model yang membedakan serta menjelaskan konsep atau *class* data, dengan tujuan untuk memprediksi *class* pada objek yang labelnya belum diketahui. Adapun metode klasifikasi yang telah dikembangkan dan sering difungsikan diantaranya seperti *Naive Bayes, C4.5, K-Nearest Neighbor,* dan *Neural Network.*

Terdapat banyak peneliti yang menggunakan metode atau algoritma klasifikasi pada *data mining* dalam memprediksi berbagai penyakit, akan tetapi masih sulit mengetahui metode atau algoritma yang paling akurat. Komparasi metode klasifikasi banyak dilakukan para peneliti dengan hasil yang berbeda pula, seperti yang dilakukan oleh Oktafia yang menyatakan bahwa dalam perbandingan unjuk kerja algoritma atau metode *Naive Bayes* dan *Decision Tree* dalam melakukan prakiraan kebangkrutan perusahaan yang menghasilkan kesimpulan bahwa algoritma *naive bayes* memiliki unjuk kerja yang lebih tinggi dengan nilai akurasi mencapai 100%. Pada penelitian ini dalam melakukan perbandingan algoritma hanya untuk kasus prediksi kebangkrutan perusahaan. Aplikasi yang digunakan yaitu WEKA dengan menggunakan tiga model pengujian yaitu *use training set, cross validation* dan *percentage split*. Penelitian ini memberi kesimpulan bahwa *cross validation* merupakan model pengujian yang lebih direkomendasikan, hal ini dikarenakan pada model tersebut, setiap data pada dataset

mempunyai peluang yang sama untuk menjadi data tes dan data latih [3].

Penelitian lainnya oleh Fatmawati dalam melakukan komparasi algoritma atau metode klasifikasi *Naive Bayes* dan C4.5 dalam memprediksi penyakit diabetes. Penelitian ini menggunakan algoritma *Decision Tree* dan menghasilkan tingkat akurasi yaitu sebesar 73.30%, sedangkan algoritma *Naive Bayes* sebesar 75.13%. Pada penelitian ini komparasi kedua algoritma tersebut hanya dilakukan untuk kasus penyakit diabetes, dalam evaluasi akurasi model klasifikasi selain menggunakan *confusion matrix* juga menggunakan kurva ROC. Keseluruhan atribut pada penelitian ini bernilai numerik, dan label *class* nya dalam bentuk binomial [4].

Pada penelitian lain yang dilakukan oleh Listiana et al. yang melakukan komparasi algoritma *Naive Bayes* dan *Decision Tree* C4.5 untuk mengidentifikasi perkembangan balita (Studi Kasus Pada Puskesmas Kartasura), hasil penelitian ini menunjukkan tingkat akurasi algoritma *Naive Bayes* sebesar 76,97% lebih unggul dibandingkan dengan algoritma C4.5 yang memiliki tingkat akurasi 75,66%. Pada penelitian ini juga melakukan perbandingan algoritma klasifikasi, akan tetapi untuk kasus identifikasi perkembangan balita. Penelitian ini menggunakan lima atribut dan satu label *class*, dimana atributnya bersifat polynomial, binomial serta numerik [5].

Maulidia et al. melakukan penelitian membangun sistem diagnosa penyakit pada anak dengan mengimplementasikan *Case Based Reasoning*, sistem ini berbasis web. Dari hasil proses pengujian pada 30 kasus, didapatkan hasil rata-rata akurasi sistem senilai 86%. Pada penelitian ini hanya membangun sebuah sistem untuk melakukan diagnosa penyakit pada anak berdasarkan kasus-kasus yang sudah pernah terjadi, bukan perbandingan algoritma [6].

Friska et al. melakukan penelitian membangun sistem pakar menggunakan metode *Dempster Shafer* untuk diagnosa penyakit pada anak. Berdasarkan proses pengujian 60 data Rekam Medis yang dilakukan pada sistem aplikasi, diperoleh hasil persentase keberhasilan sebesar 88,33%. Pada penelitian ini hanya membangun sebuah sistem pakar dengan penerapan metode *Dempster Shafer* untuk mendiagnosa penyakit pada anak, bukan perbandingan algoritma [7].

Berdasarkan penjelasan di atas, pada penelitian ini akan dilakukan proses perbandingan algoritma atau metode *Decision Tree* C4.5 dan *Naive Bayes* untuk klasifikasi penyakit anak. Pemilihan penggunaan kedua algoritma tersebut lebih dikarenakan merupakan algoritma yang sangat populer dan banyak digunakan secara praktis. Penelitian ini dilakukan agar dapat diketahui algoritma yang memiliki nilai akurasi tertinggi dalam mendeteksi penyakit pada anak.

II. LANDASAN TEORI

A. Data Mining

Data mining merupakan proses analisa dan penggalian data yang besar agar dapat diperoleh suatu kebenaran, hal yang baru serta memiliki manfaat hingga dapat ditemukan pola tertentu pada data tersebut [8].

Proses penggalian data tersebut dilakukan secara iterasi atau berulang-ulang hingga didapatkan hasil satu set pola yang sesuai dan berfungsi sesuai tujuan awal. *Data mining* secara garis besar dikelompokkan menjadi dua kategori [9] :

1. *Predictive mining* merupakan proses mendapatkan suatu pola pada data menggunakan beberapa variabel untuk memprediksi variabel lainnya pada waktu yang akan datang. Teknik yang termasuk dalam *predictive* diantaranya seperti regresi, deviasi dan klasifikasi.
2. *Descriptive mining* merupakan proses untuk mendapatkan karakteristik penting pada data dalam suatu *database*. Teknik yang termasuk dalam *descriptive mining* diantaranya seperti *sequential mining*, *association* dan *clustering*.

Data mining merupakan bagian integral dari penemuan pengetahuan dalam basisdata atau dikenal dengan *Knowledge Discovery in Databases* (KDD). KDD merupakan semua proses pengubahan data mentah menjadi pola yang menarik dan merupakan pengetahuan berupa bagian informasi yang dibutuhkan oleh pengguna.

B. Klasifikasi

Klasifikasi adalah proses menemukan suatu model atau pola yang membedakan serta menggambarkan konsep atau *class* data, tujuannya adalah agar model tersebut dapat digunakan untuk memprediksi *class* objek yang label *class* nya belum diketahui. Model didapatkan berdasarkan pada analisis pada serangkaian data pelatihan (yaitu, objek data yang sudah diketahui label kelasnya) [8].

Pengelompokan data tersebut akan mempelajari data sampel menggunakan algoritma klasifikasi dengan mengenali pola tertentu pada data sampel terhadap kelas target, sehingga nantinya memungkinkan untuk dapat melakukan prediksi kelas target dengan menggunakan data diluar data sampel dengan menggunakan model *classifier* [8]. Contoh dalam dunia kedokteran, untuk menentukan bahwa seorang pasien divonis positif atau tidak terhadap suatu penyakit, maka dilakukan proses *learning* menggunakan data-data hasil pemeriksaan dan gejala yang dialami oleh pasien beserta hasil penyakit. Langkah selanjutnya dengan menggunakan model *classifier* dari hasil *learning* tersebut vonis penyakit bisa diprediksi.

Menurut Han dan Kamber dalam klasifikasi terdapat dua tahapan proses. Pada tahap pertama model *classifier* akan dibentuk dengan cara menganalisis data set atau data *training* menggunakan algoritma klasifikasi. Proses ini disebut dengan tahap pembelajaran (*learning step*) atau fase *training* [8].

Pada tahap kedua dilakukan evaluasi terhadap model *classifier* untuk mendapatkan nilai akurasi. Akurasi dari model klasifikasi yang diberikan oleh data *test* adalah persentase data *test* yang diklasifikasikan dengan benar oleh *classifier*. Jika nilai akurasi sesuai dengan yang diharapkan, maka *classifier* sudah dapat digunakan dan dilakukan proses klasifikasi. Proses klasifikasi tersebut dilakukan untuk mendapatkan hasil prediksi menggunakan tupel data diluar data *training*. Model atau *rule* yang telah didapatkan selanjutnya digunakan untuk proses klasifikasi pada data yang belum diketahui (*unknown*). Akan tetapi, hanya model yang memiliki akurasi yang cukup tinggi yang dapat digunakan untuk proses klasifikasi.

Akurasi didapatkan dengan cara model atau *rule* yang dihasilkan dari proses *training* diuji dengan data *test*. Data t*est* terdiri dari *record-record* yang label *class* nya sudah diketahui, akan tetapi data *test* tidak boleh sama dengan data *training*, hal ini akan membuat proses pengujian menghasilkan akurasi yang tinggi, padahal belum tentu seperti itu.

C. Algoritma Naive Bayes

Algoritma *naive bayes* atau sering disebut *Naive bayes classifier* (NBC) merupakan salah satu algoritma pada metode klasifikasi yang dapat memprediksi probabilitas atau kemungkinan keanggotaan pada suatu *class*. NBC mengasumsikan nilai atribut pada suatu *class* bersifat independen atau bebas terhadap nilai pada atribut yang lain.

Algoritma *Naive Bayes* memiliki tahapan sebagai berikut [8] :

Setiap contoh data diwakilkan dengan *n*-dimensional *feature vector*, $X=(X_1,X_2,X_3\ldots,X_n)$, dimana *n* dibuat dari contoh *n* atribut, berturut-turut $A_1,A_2,A_3\ldots,A_n$.

Diandaikan ada *m class*, $C_1,C_2,C_3\ldots,C_m$. Diberikan sebuah contoh data, *X* (yang belum diketahui label *class* nya), selanjutnya *classifier* akan melakukan prediksi *X* ke dalam *class* yang memiliki nilai probabilitas posterior paling tinggi berdasarkan kondisi *X*. Naive bayes classifier akan menetapkan bahwa sample data *X* yang belum diketahui sebelumnya ke dalam *class* C_i hanya jika :

$P(C_i|X) > P(C_j|X)$ *untuk* $1 \leq j \leq m$, dimana $j \neq i$

Berdasarkan kondisi tersebut, $P(C_i \mid X)$ harus dimaksimalkan. Pada $P(C_i \mid X)$ *Class* C_i merupakan nilai terbesar atau disebut maksimum posteri hypothesis,

$$P(C_i \mid X) = \frac{P(X \mid Ci)\, P(Ci)}{P(X)} \quad (1)$$

Nilai $P(X)$ tetap untuk semua *class*, yang perlu dimaksimalkan hanya $P(X|C_i)P(C_i)$. Jika probabilitas *prior class* belum diketahui, umumnya ditetapkan nilai setiap *class* sama, yaitu $P(C_1)=P(C_2)=P(C_3)\ldots=P(C_m)$, dan berikutnya yaitu hitung nilai $P(X \mid C_i)$ dan hitung nilai $P(X \mid C_i)\, P(C_i)$. Probabilitas *class prior* dapat dihitung dengan cara :

$$P(\mathrm{C_i}) = \frac{s_i}{s} \quad (2)$$

s_i merupakan jumlah *training* sample pada *class Ci* di dalam data *training*, sedangkan *s* merupakan jumlah total data *training sample*.

Jika dataset mencakup banyak atribut, hal ini dapat membuat proses komputasi akan rumit untuk mengestimasi $P(X|C_i)$. Dalam rangka pengurangan beban komputasi dalam mengevaluasi $P(X|C_i)$, *naive bayes* menetapkan pembuatan *class* bersifat independen. Jadi nilai atribut diatur agar bersifat bebas atau independen antara atribut satu dengan atribut lainnya, serta tidak terdapat relasi depedensi antar atribut.

$$P(\mathrm{X} \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i) \quad (3)$$

$$= P(x_1|C_i)\; x\; P(x_2|C_i)\; x \ldots x\; P(x_n|C_i)$$

Proses estimasi probabilitas $P(x_1|C_i)$, $P(x_2|C_i)$, $P(x_3|C_i) \ldots$, $P(x_n|C_i)$ dari data *training* akan lebih mudah dilakukan. Dimana x_k adalah nilai pada atribut A_k pada data *X*. Pada tiap atribut, dilihat atribut tersebut apakah bernilai kontinyu atau kategorikal, sehingga untuk melakukan perhitungan $P(X|C_i)$ mengikuti aturan seperti berikut ini :

1. Jika nilai A_k bersifat kategorikal, perhitungan $P(X|C_i)$ menjadi

$$P(x_k \mid C_i) = \frac{s_{ik}}{s_i} \quad (4)$$

s_{ik} merupakan jumlah *training sample* pada *class* C_i yang memiliki nilai x_k untuk A_k sedangkan s_i merupakan jumlah *training sample* yang masuk ke dalam *class* C_i.

2. Jika nilai A_k bersifat kontinyu, terdapat sedikit perbedaan. Sebuah atribut bernilai kontinyu umumnya memiliki distribusi Gaussian yang memiliki *mean* (rata-rata) μ serta standar deviasi σ, yang didefinisikan dengan

$$g(x_k, \mu C_i, \sigma C_i) = \frac{1}{\sqrt{2\pi}\,\sigma C_i}\, e^{-\frac{(x_k-\mu C_i)^2}{2\,\sigma^2 C_i}} \quad (5)$$

sehingga

$$P(X_k \mid C_i) = g(x_k, \mu C_i, \sigma C_i) \quad (6)$$

dimana,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (7)$$

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2 \quad (8)$$

Persamaan (6) merupakan fungsi gaussian pada atribut A_k dengan μC_i serta σC_i masing-masing merupakan *mean* (rata-rata) dan standar deviasi dari nilai atribut A_k pada training sample *class* C_i.

Proses estimasi label *class* pada sample *X*, $P(X \mid C_i)\, P(C_i)$ dievaluasi setiap *class* C_i. *Classifier* mengestimasi label *class* pada *sample X* merupakan *class* C_i hanya jika

$P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$ *untuk* $1 \leq j \leq m, j \neq i$ (9)

Persamaan (9) menunjukkan bahwa hasil perkiraan label *class* merupakan *class* C_i pada $P(X \mid C_i)P(C_i)$ yang memiliki nilai maksimum.

D. Algoritma C4.5

Secara umum, *Decision Tree* (pohon keputusan) merupakan suatu bentuk model dari suatu masalah yang tersusun dari rangkaian keputusan yang menunjuk pada pemecahan masalah yang dihasilkan [8].

Pembentukan pohon keputusan (*Decision Tree*) dapat menggunakan Algoritma C4.5. Pohon keputusan dimanfaatkan untuk meneliti data lebih lanjut untuk menemukan relasi yang masih tersembunyi antara beberapa calon variabel *input* dan sebuah variabel *output* atau target.

Pohon keputusan dapat dipandang sebagai suatu pendekatan yang paling terkenal. Pohon keputusan tersusun atas sebuah simpul yang digambarkan sebagai akar, simpul akar tersebut

tidak memiliki *input*. Selain itu terdapat simpul *internal* atau simpul *test* yang merupakan simpul lain yang bukan sebagai akar, akan tetapi memiliki tepat hanya satu masukan, dan simpul lainnya disebut daun. Daun merupakan representasi nilai *output* atau target dari salah satu *class* yang paling tepat [10].

Berikut tahapan algoritma *Decision Tree* C4.5 secara umum dalam membangun pohon keputusan [11] :

1. Akar dipilih dari salah satu atribut.
2. Tiap nilai dibuat cabang.
3. Pada cabang dibagi kasus.
4. Hingga setiap kasus memiliki kelas yang sama pada cabang maka ulangi proses tersebut.

Pemilihan akar dari atribut, berdasarkan pada atribut yang memiliki nilai *gain* tertinggi. Persamaan (10) digunakan untuk menghitung *gain* :

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\,(S_i) \quad (10)$$

Keterangan :
S : kelompok kasus
A : atribut
n : banyaknya bagian atribut *A*
$|S_i|$: banyaknya kasus pada bagian ke-*i*
|S| : banyaknya kasus dalam *S*

Persamaan (11) digunakan untuk menghitung nilai *entropi* :

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\, pi \quad (11)$$

Keterangan :

S : kelompok kasus.
n : banyaknya partisi *S*.
pi : perbandingan dari S_i terhadap *S*

E. Cross Validation

Metode *Cross Validation* merupakan metode yang digunakan untuk mengevaluasi kinerja model maupun algoritma. Salah satu variasi dari teknik pengujian *cross validation* yaitu *k-fold*. Metode ini dilakukan dengan membagi sampel data menjadi *k* bagian yang rata untuk digunakan sebagai data *training set* dan data *test set*, secara berulang-ulang *k* kali.

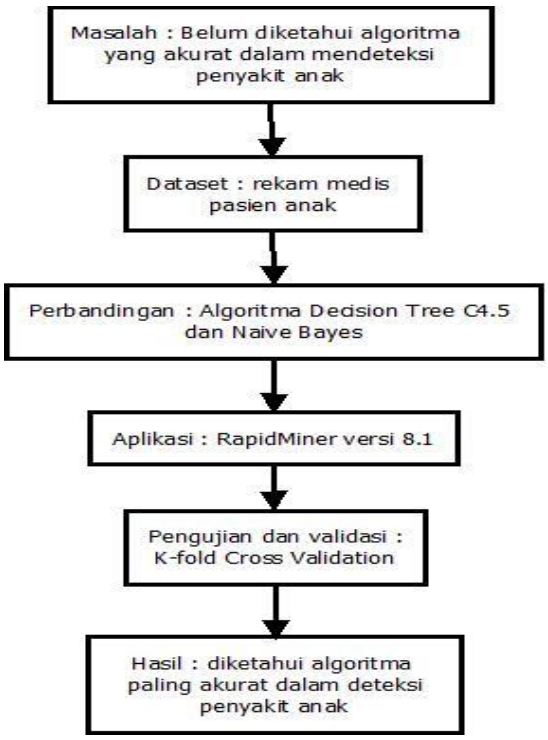
Secara prinsip, inti dari metode validasi ini adalah membagi data secara acak ke dalam *k* subset (himpunan bagian) yang saling bebas. Pada *k* himpunan bagian tersebut sebanyak (k-1) menjadi data latih dan satu himpunan bagian *k* menjadi data uji. Perhitungan estimasi ketelitian atau akurasi dari metode ini adalah jumlah keseluruhan klasifikasi yang benar dibagi dengan jumlah kasus dalam dataset [12]. Menghitung nilai akurasi akhir model didapat dengan merata-ratakan nilai akurasi untuk *k* percobaan yang dilakukan, ditunjukkan pada Persamaan (12). Model hasil pengujian yang akan digunakan adalah model yang nilai akurasinya paling tinggi diantara model yang ada.

$$Akurasi = \frac{Jumlah\ Klasifikasi\ Benar}{Jumlah\ Data\ Uji} \times 100\% \quad (12)$$

III. METODOLOGI PENELITIAN

A. Kerangka Pemikiran

Penelitian ini merupakan bentuk penelitian model eksperimen menggunakan alat bantu *Tools* RapidMiner versi 8.1. Tujuan penelitian ini adalah untuk membandingkan akurasi dari tiap algoritma agar diketahui algoritma terbaik dalam mengklasifikasi penyakit anak. Algoritma yang akan dilakukan perbandingan yaitu Naive Bayes dan Decision Tree C4.5. Pengujian dan validasi menggunakan *k-fold Cross Validation.* Kerangka pemikiran dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Kerangka Pemikiran

Penelitian ini dilakukan beberapa tahapan mulai dari pengumpulan data untuk dijadikan dataset penelitian, pengolahan data penelitian, pengukuran hasil pengujian, dan analisa perbandingan hasil pengujian.

B. Dataset

Penelitian ini menggunakan dataset berupa data rekam medis pasien anak dari Rumah Sakit Islam Yarsi Pontianak sebanyak 240 *record* (kasus) yang terjadi selama tahun 2014 hingga 2016. Dataset tersebut terdiri dari 52 atribut berupa gejala penyakit anak dengan nilai masing-masing berupa teks ‘Ya’ atau ‘Tidak’. Serta 1 atribut label yang memiliki nilai berupa 12 kelas berupa penyakit pada anak. Tabel I menunjukkan potongan dataset rekam medis penyakit anak.

TABEL I. DATASET REKAM MEDIS PASIEN ANAK

NO	Umur	Jenis kelamin	Hasil diagnosa penyakit	Demam tinggi mendadak dan berlangsung 2-7 hari (38-40° Celsius)	Batuk berdahak	Pilek	Dada anak terasa sakit	Setiap diberikan minum muntah	...
1	Z(10 thn)	perempuan	Demam berdarah dengue atau DHF	Ya	Tidak	Tidak	Tidak	Tidak	...
2	SY(10 thn)	perempuan	Demam tifoid (tifus)	Tidak	Tidak	Tidak	Tidak	Tidak	...
3	HY(10 thn)	Laki-laki	ISPA (Infeksi saluran pernapasan akut)	Ya	Ya	Ya	Tidak	Tidak	...
4	J (5 thn)	Perempuan	Varicella (cacar air)	Ya	Tidak	Tidak	Tidak	Tidak	...
5	TL(12thn)	perempuan	Tuberkolosis (TB)	Ya	Tidak	Tidak	Tidak	Tidak	...
6	C(7 thn)	Laki-laki	KDK (kejang demam kompleks)	Ya	Tidak	Tidak	Tidak	Tidak	...
7	M(8 thn)	Perempuan	Vomitus	Tidak	Ya	Ya	Tidak	Ya	...
...	...	...	...	...	...	...	...	...	...

Data rekam medis pasien anak tersebut diolah dan dapat dirincikan berupa 12 kelas penyakit yang dapat dilihat pada Tabel II dan 52 gejala penyakit anak sebagai atribut yang dapat dilihat pada Tabel III.

TABEL II. DATA PENYAKIT ANAK

No	Nama Penyakit
1	Anemia Defisiensi Besi (ADB)
2	Asma Bronchial
3	Demam Berdarah Dengue (DBD)
4	Demam Tifoid (Tipes)
5	Diare
6	Diare Akut
7	ISPA (Infeksi Saluran Pernapasan Akut)
8	Kejang Demam Komplek (KDK)
9	Kejang Demam Sederhana (KDS)
10	Tuberkolosis Paru (TB-paru)
11	Varisella (Cacar air)
12	Vomitus (Muntah)

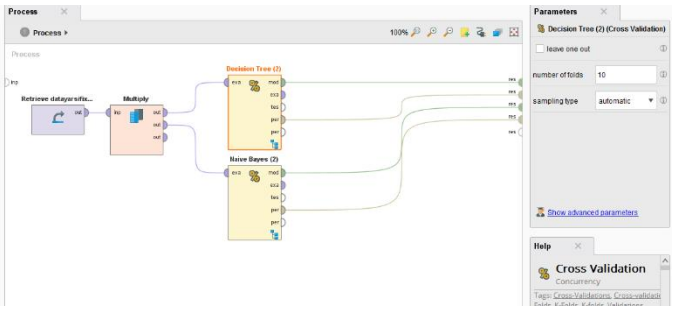
TABEL III. DATA GEJALA PENYAKIT ANAK

No	Nama Gejala
1	Demam tinggi mendadak dan berlangsung 2-7 hari (38-40° Celsius)
2	Wajah anak terlihat pucat dalam waktu tertentu
3	Telapak tangan anak terlihat pucat
4	Terasa mengambang (gelap) pada saat bangun tidur dan saat berdiri
5	Sakit kepala (pusing)
6	Badan anak terlihat lemas/lesu
7	Tidak napsu makan
8	Mengalami Mual
9	Mengalami Muntah
10	Nyeri perut (sakit perut)
11	Batuk kering
12	Batuk berdahak
13	Pilek
14	Mengalami sesak napas kuat
15	Mempunyai riwayat asma sebelumnya
16	Dada anak terasa sakit
17	Mengalami susah tidur (gelisah saat tidur)
18	Timbul bintik-bintik (ruam) kemerahan dikulit tubuh
19	Mengalami mimisan (pendarahan dihidung)
20	Mengalami pendarahan digusi (gusi berdarah)
21	Demam naik-turun lebih dari 7 hari (39-40° Celsius)
22	Badan anak terasa mengigil
23	Mengalami perubahan pola buang air besar (Susah BAB) dalam waktu tertentu
24	Buang air besar (BAB) terus-menerus selama lebih dari 3x sehari.
25	Ada bercak darah ditinja (kotoran)
26	Ada lendir ditinja (kotoran)
27	Buang air kecil (BAK) berkurang
28	Buang air besar (BAB) terus-menerus dan lebih encer (cair) dari biasanya selama kurang dari 14 hari
29	Batuk berdahak
30	Mengalami sakit tenggorokan (Susah menelan)
31	Kelonjotan/ kejang diseluruh tubuh lebih dari 15 menit
32	Mengalami kejang 2X dalam waktu lebih dari 15 menit dalam waktu 24 jam
33	Gerak bola mata anak terlihat mendelik keatas (juling keatas)
34	Ada riwayat kejang sebelumnya
35	Ada menangis setelah mengalami kejang
36	Kelonjotan/kejang diseluruh tubuh kurang dari 15 menit
37	Mengalami kejang 1X dalam waktu kurang dari 15 menit dalam waktu 24 jam
38	Mengalami kejang untuk pertama kali
39	Mengalami batuk lama lebih dari 3 minggu
40	Ada terlihat/teraba benjolan dibagian leher
41	Mengalami Kesulitan dalam berat badan (berat badan turun selama minimal 1 bulan)
42	Saat batuk ada bercampur dengan darah (mengeluarkan darah)
43	Mengalami berkeringat dingin (pada malam hari)
44	Mengalami suara serak
45	Muncul bintil-bintil diseluruh tubuh (bintil muncul mulai dari badan lalu ke tangan dan kaki)
46	Bintil-bintil (bentolan kecil) berisikan cairan (nanah)
47	Bintil-bintil (bentolan kecil) terasa gatal
48	Bintil-bintil (bentolan kecil) terasa nyeri/perih
49	Mengalami muntah lebih sering dengan frekuensi lebih dari 3X / perhari
50	Setiap diberikan makan muntah
51	Setiap diberikan minum muntah
52	Anak terlihat gelisah (lebih cerewet).

IV. HASIL DAN PEMBAHASAN

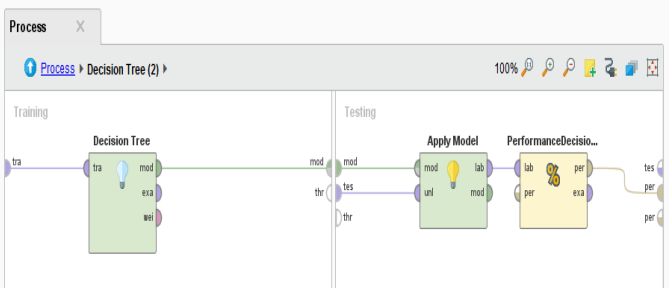
A. *Penerapan Algoritma Klasifikasi*

Berdasarkan dataset yang telah diolah, tahapan berikutnya adalah melakukan proses perhitungan tingkat akurasi untuk setiap algoritma atau metode baik Naive Bayes maupun Decision Tree C4.5 menggunakan aplikasi RapidMiner versi 8.1. Gambar 2 merupakan desain model proses utama perbandingan algoritma untuk klasifikasi data penyakit anak.



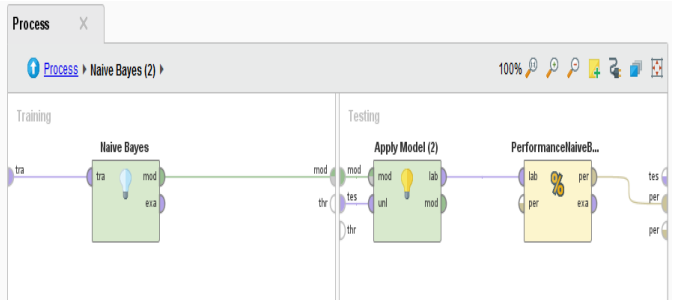
Gambar 2. Desain Model Perbandingan Algoritma Klasifikasi

Desain model proses utama ini langsung menggunakan satu dataset yang sama untuk dua algoritma sekaligus yaitu Decision Tree C4.5 serta Naive Bayes untuk memastikan konsistensi dataset yang sama yang diproses oleh setiap algoritma. Gambar 3 merupakan rincian desain model proses *training* dan *testing* untuk proses klasifikasi algoritma Decision Tree C4.5.



Gambar 3. Desain Model Training dan Testing Algoritma C4.5

Gambar 4 merupakan desain model proses *training* dan *testing* untuk proses klasifikasi algoritma Naive Bayes.



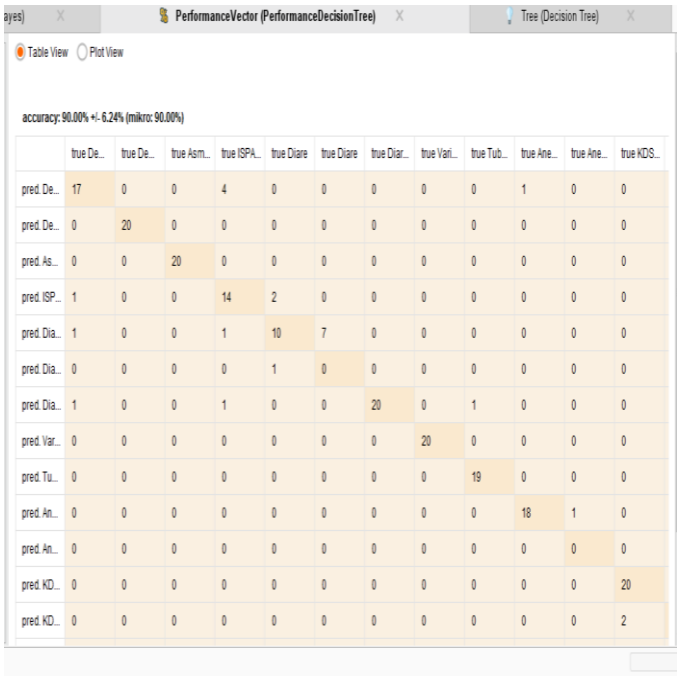
Gambar 4. Desain Model Training dan Testing Algoritma Naive Bayes

B. Pengukuran Hasil Perbandingan Algoritma

Metode *k-fold Cross Validation* digunakan pada proses pengujian dengan nilai *k* yaitu 10. Proses tersebut membagi dataset menjadi 10 dimana 1 bagian sebagai data *testing* dan 9 bagian sebagai data *training*. Proses ini berulang (iterasi) sebanyak nilai *k* yaitu 10 kali hingga seluruh bagian tersebut secara bergantian mendapat giliran sebagai data *testing*. Berdasarkan hasil pengujian tersebut didapatkan hasil akurasi dari setiap algoritma klasifikasi.

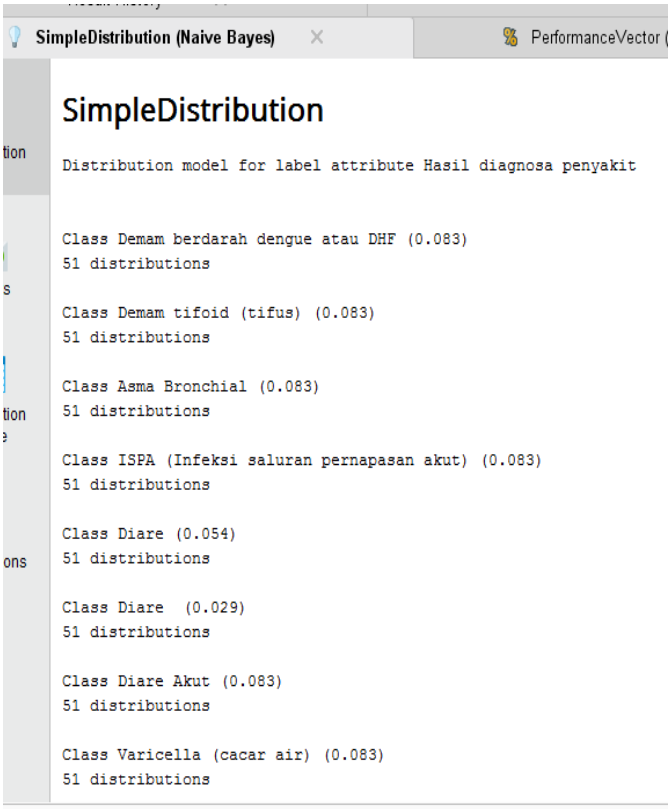
Gambar 5 merupakan pohon keputusan yang dihasilkan oleh proses klasifikasi menggunakan algoritma Decision Tree C4.5 serta Gambar 6 merupakan tingkat akurasi yang dihasilkan dari proses klasifikasi tersebut.

Gambar 5. Pohon Keputusan Algoritma C4.5

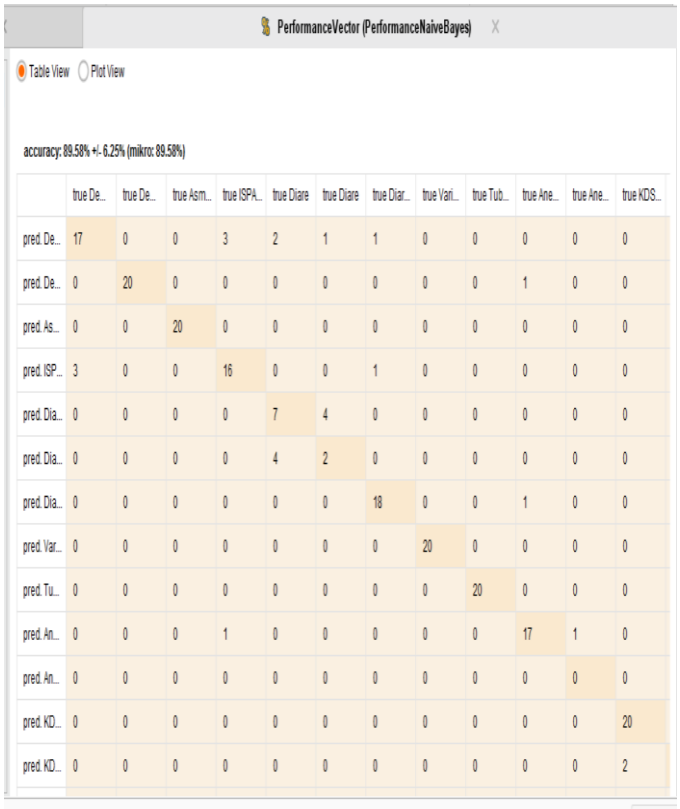


Gambar 6. Tingkat Akurasi Algoritma C4.5

Algoritma Naive Bayes menghasilkan *Simple Distribution* yang potongannya terlihat pada Gambar 7 serta Gambar 8 berupa hasil tingkat akurasi.



Gambar 7. Simple Distribution Algoritma Naive Bayes



Gambar 8. Tingkat Akurasi Algoritma Naive Bayes

Hasil tingkat akurasi yang didapatkan dari kedua algoritma, selanjutnya dilakukan pembandingan dari kedua nilai akurasi tersebut. Tabel IV merupakan hasil perbandingan tingkat akurasi dari algoritma Naive Bayes serta Decission Tree C4.5.

TABEL IV. HASIL PERBANDINGAN ALGORITMA NAIVE BAYES DAN C4.5

No	Algoritma	Tingkat Akurasi (%)
1	Naive Bayes	89.58
2	Decision Tree C4.5	90.00

Hasil perbandingan tingkat akurasi tersebut menunjukkan bahwa dalam melakukasi klasifikasi penyakit pada anak, nilai tingkat akurasi algoritma Decision Tree C4.5 lebih tinggi dibanding nilai tingkat akurasi Naive Bayes.

Hasil perbandingan ini tentu berbeda dengan hasil penelitian yang dilakukan oleh Fatmawati dan penelitian Listiana et al. dimana pada kedua penelitian tersebut menunjukkan bahwa kinerja algoritma Naive Bayes memiliki tingkat akurasi yang lebih tinggi dibanding algoritma *Decision Tree*. Hal ini selain studi kasus yang berbeda, karakteristik dari atribut yang diolah juga berbeda. Pada penelitian yang dilakukan memiliki jumlah atribut yang lebih banyak dengan hanya dua buah nilai atribut, sementara pada dua penelitian sebelumnya tersebut memiliki jumlah atribut yang lebih sedikit tapi dengan banyak nilai atribut.

V. KESIMPULAN

Penelitian ini menggunakan dataset rekam medis pasien anak, dimana atribut yang diolah oleh algoritma klasifikasi berupa gejala-gejala penyakit pada anak dengan nilai ‘ya’ atau ‘tidak’ untuk setiap atribut gejalanya. Serta terdapat sebuah label *class* dengan nilai berupa jenis-jenis penyakit yang diderita oleh anak.

Pada penelitian ini analisis yang dilakukan menggunakan aplikasi RapidMiner versi 8.1 serta metode *K-fold Cross Validation* untuk menguji tingkat akurasi, dan didapatkan hasil bahwa algoritma *Decision Tree* C4.5 merupakan algoritma terbaik dalam melakukan klasifikasi penyakit anak dengan tingkat akurasi sebesar 90%, sedangkan algoritma Naive Bayes memperoleh tingkat akurasi 89.58%.

REFERENSI

[1] M. Septa, H. Salwan, and S. Tjekyan, “Pengaruh Suplementasi Vitamin A Terhadap Lama Diare pada Anak Usia 12-51 Bulan yangBerobat di Puskesmas Sukarami Palembang,” *J. Kedokt. dan Kesehat.*, vol. 2, no. 2, pp. 117–123, 2015.

[2] M. Neshat and M. Yaghobi, “Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network Fuzzy System,” in *Proceedings of the World Congress on Engineering and Computer Science*, 2009, vol. II.

[3] D. Oktafia, “Perbandingan Kinerja Algoritma Decision Tree dan Naive Bayes dalam Prediksi Kebangkrutan,” *J. Ilm. Ilmu Komput. Progr. Stud. Sist. Inf. (Universitas Gunadarma)*, 2007.

[4] Fatmawati, “Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 dan Naive Bayes untuk Prediksi Penyakit Diabetes,” *J. Techno Nusa Mandiri*, vol. XIII, no. 1, pp. 50–59, 2016.

[5] M. Listiana, Sudjalwo, and D. Gunawan, “Perbandingan Algoritma Decision Tree (C4.5) Dan Naïve Bayes Pada Data Mining Untuk Identifikasi Tumbuh Kembang Anak Balita (Studi Kasus Puskesmas Kartasura),” *Informatika*, vol. 1, no. 1, p. 18, 2015.

[6] T. R. Maulidia, T. Rismawan, and S. Bahri, “Implementasi Case Based Reasoning Sistem Diagnosa Penyakit Anak Berbasis Web,” *J. Coding, Sist. Komput. Untan*, vol. 5, no. 3, pp. 57–63, 2017.

[7] A. I. Friska, T. Rismawan, and S. Bahri, “Aplikasi Sistem Pakar Diagnosa Penyakit Pada Anak Dengan Inference Forward Menerapkan Metode Dempster Shafer Berbasis Web,” *J. Coding, Sist. Komput. Untan*, vol. 06, no. 02, pp. 25–35, 2018.

[8] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 54, no. Second Edition. San Francisco: Morgan Kaufmann, 2006.

[9] P.-N. Tan, M. Steinbach, and V. Kumar, “Introduction to Data Mining,” Pearson Addison-Wesley, 2006, p. 169.

[10] L. Maimon, o., & Rokach, *Data Mining and Knowladge Discovery Handbook Second Edition*. New York: Springer., 2010.

[11] T. E. Kusrini, & Luthfi, *Algoritma Data Mining*. Yogyakarta: Penerbit Andi, 2009.

[12] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *Appear. Int. Jt. Conf. Articial Intell.*, vol. 5, pp. 1–7, 1995.