

## Analisis Komparatif Evaluasi Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes

Mochammad Yusa<sup>1</sup>, Ema Utami<sup>2</sup>, Emha T. Luthfi<sup>3</sup>

Magister Teknik Informatika, STMIK AMIKOM Yogyakarta

Jl. Ring Road Utara, Condong Catur, Sleman, 55283, DI Yogyakarta, Indonesia

E-mail: <sup>1</sup>mochammad.yusa@amikom.ac.id, <sup>2</sup>ema.u@amikom.ac.id, <sup>3</sup>emhataufiqluthfi@amikom.ac.id

Masuk: 12 Mei 2016; Direvisi: 6 Juni 2016; Diterima: 9 Juni 2016

**Abstract.** Readmission is associated with quality measures on patients in hospitals. Different attributes related to diabetic patients such as medication, ethnicity, race, lifestyle, age, and others result in the calculation of quality care that tends to be complicated. Classification techniques of data mining can solve this problem. In this paper, the evaluation on three different classifiers, i.e. Decision Tree, *k*-Nearest Neighbor (*k*-NN), dan Naive Bayes with various setting parameter, is developed by using 10-Fold Cross Validation technique. The targets of parameter performance evaluated is based on term of Accuracy, Mean Absolute Error (MAE), dan Kappa Statistic. The selected dataset consists of 47 attributes and 49.735 records. The result shows that *k*-NN classifier with *k*=100 has a better performance in term of accuracy and Kappa Statistic, but Naive Bayes outperforms in term of MAE among other classifiers.

**Keywords:** *k*-NN, naive bayes, diabetes, readmission

**Abstrak.** Proses Readmisi dikaitkan dengan perhitungan kualitas penanganan pasien di rumah sakit. Perbedaan atribut-atribut yang berhubungan dengan pasien diabetes proses medikasi, etnis, ras, gaya hidup, umur, dan lain-lain, mengakibatkan perhitungan kualitas cenderung rumit. Teknik klasifikasi data mining dapat menjadi solusi dalam perhitungan kualitas ini. Teknik klasifikasi merupakan salah satu teknik data mining yang perkembangannya cukup signifikan. Di dalam penelitian ini, model algoritma klasifikasi Decision Tree, *k*-Nearest Neighbor (*k*-NN), dan Naive Bayes dengan berbagai parameter setting akan dievaluasi performanya berdasarkan nilai performa Accuracy, Mean Absolute Error (MAE), dan Kappa Statistik dengan metode 10-Fold Cross Validation. Dataset yang dievaluasi memiliki 47 atribut dengan 49.735 records. Hasil penelitian menunjukkan bahwa performa accuracy, MAE, dan Kappa Statistik terbaik didapatkan dari Model Algoritma Naive Bayes.

**Kata Kunci:** *k*-NN, naive bayes, diabetes, readmisi

### 1. Pendahuluan

Readmisi Rumah Sakit (*Hospital Readmission*) adalah suatu tindakan atau kejadian seorang pasien dirawat kembali yang sebelumnya telah mendapatkan layanan rawat inap dirumah sakit (Iskandar, 2014). Proses readmisi dikaitkan dengan perhitungan kualitas penanganan pasien di rumah sakit (Rubin, 2015; Dungan, 2012). Penyakit diabetes merupakan penyakit metabolisme tubuh yang saat ini belum ada obat yang dapat menyembuhkannya dan dapat menimbulkan penyakit komplikasi-komplikasi lainnya. Pasien penderita diabetes dituntut untuk menjaga tingkat glukosa dalam tubuh agar tetap berada di level yang sehat. Karena tingkat perawatan inap pasien diabetes berada di angka yang tinggi (Ramirez, dkk., 2012), sehingga biaya perawatan yang dibayarkan cenderung mahal (American Diabetes Association, 2013).

Perhitungan kualitas terhadap pelayanan penyakit diabetes dapat dilihat dari proses medikasi, etnis, ras, gaya hidup, umur, dll. Di dalam penyakit diabetes, perbedaan atribut terkait pasiennya sangat berpengaruh terhadap kualitas pengobatan yang merujuk pada ketahanan

serum glikemik yang ada di dalam tubuh. Semakin lama serum glikemik berada di tingkat yang sehat, maka semakin baik kualitas pengobatan atau medikasi rumah sakit tersebut (Dungan, 2012). Namun perbedaan atribut-atribut yang berkaitan dengan pasien diabetes mengakibatkan perhitungan kualitas cenderung rumit. Teknik klasifikasi *data mining* dapat menjadi solusi dalam perhitungan kualitas ini. Teknik klasifikasi merupakan salah satu teknik *data mining* yang perkembangannya cukup signifikan.

Dalam teknik klasifikasi *data mining* terdapat banyak model-model algoritma klasifikasi dengan nilai performa yang berbeda-beda. Rahman dan Afroz (2013) menyatakan bahwa model algoritma yang baik pada satu kasus belum tentu algoritma tersebut mempunyai nilai yang baik pada kasus lain. Hal ini diperkuat dengan penelitian-penelitian di Kesehatan (Temurtas, dkk., 2009, Tomar & Agarwal, 2013; Kaur & Singh, 2014) yang berkaitan tentang pengimplementasian teknik klasifikasi pada *dataset-dataset* kesehatan baik berupa diagnosis maupun analisis medikasi.

Data readmisi pasien diabetes berisikan data dengan atribut dan *record* yang sangat banyak dan bersifat heterogen, sehingga teknik klasifikasi dibutuhkan guna mencari solusi dalam penentuan algoritma yang memiliki performa terbaik perlu dilakukan agar nantinya dapat digunakan untuk perhitungan pengukuran terkait seberapa baik penanganan medis yang telah dilakukan terkait penyakit diabetes. Selain itu pengukuran performa algoritma dapat juga dijadikan landasan seberapa optimal performa model-model algoritma tersebut dalam mengklasifikasikan data dengan jumlah *record* dan atribut yang cenderung besar dan heterogen.

Berdasarkan fenomena-fenomena tersebut, evaluasi performa algoritma klasifikasi sangat berpengaruh terhadap hasil yang akan didapatkan terkait proses perhitungan kualitas terkait readmisi pasien diabetes, sehingga evaluasi model algoritma klasifikasi pada *dataset* tertentu untuk mendapatkan hasil performa yang optimal akan menjadi sangat krusial. Kemudian di dalam dunia kesehatan, Yoon, dkk. (2013) menyatakan bahwa kebenaran nilai diagnosis terhadap penyakit pasien sangat bergantung pada performa atau *knowledge* yang diterapkan pada sistem atau *tool* yang digunakan.

## 2. Tinjauan Pustaka

### 2.1. Penelitian Terdahulu

Pendekatan-pendekatan tentang perbandingan model-model algoritma klasifikasi *Decision Tree*, *k-NN*, dan *Naive Bayes* terkait pencarian performa yang terbaik dari berbagai bidang telah banyak dilakukan oleh peneliti-peneliti lain. Berikut ini adalah literatur *review* yang telah di-*review* sebagai landasan penelitian ini.

Durairaj & Deepika (2015) membandingkan tiga tradisional model algoritma klasifikasi seperti *Naive Bayes*, *k-NN* (*lazy classifiers*), dan *Decision Tree* berdasarkan nilai performa akurasi dan *time execution* pada *dataset* kanker leukimia yang *datasetnya* terdiri dari 7.130 atribut dan 72 *records*. Penelitian ini membuktikan pada algoritma *Naive Bayes* memiliki nilai performa akurasi yang terbaik yaitu 91,17% daripada model algoritma klasifikasi lainnya yaitu *Decision Tree* dan *k-NN*. Shaikh, dkk. (2015) mengevaluasi performa dari tiga model algoritma klasifikasi yaitu *k-NN*, *Naive Bayes*, dan *Decision Tree* pada *dataset* medis pasien penyakit jantung. Penelitian membagi percobaan menjadi tiga kasus yaitu kasus pertama dengan percobaan menggunakan semua atribut, kasus kedua menggunakan delapan atribut, dan kasus yang ketiga menggunakan tujuh atribut. Atribut tersebut digunakan sebagai acuan dasar dalam percobaan model algoritma klasifikasi.

Kemudian penelitian Fitri (2014) yang menunjukkan bahwa model algoritma klasifikasi *Naive Bayes* memiliki nilai akurasi yang baik setelah dievaluasi pada dua *dataset* berbeda yaitu *Ecoli* (delapan atribut dan 336 *records*) dan *Yeast* (sembilan atribut dan 1.484 *records*). Ashari, dkk. (2013) mencari algoritma terbaik sebagai upaya perancangan alternatif pada sebuah alat simulasi energi pada bangunan dengan mengevaluasi tiga algoritma terpopuler yaitu *Decision Tree*, *Naive Bayes*, dan *k-NN* untuk diukur nilai performanya sebagai landasan algoritma dalam perancangan alat simulasi energi pada bangunan tersebut. Karim dan Rahman (2013) membandingkan performa dua model algoritma klasifikasi yaitu *Naive Bayes* dan *C4.5 decision*

*tree* (*Gain Ratio*) untuk memprediksi apakah klien tersebut akan menabung atau tidak menabung. Kesimpulan dari penelitian ini menunjukkan bahwa algoritma *Decision Tree C4.5* memiliki nilai akurasi sebesar 94% sedangkan *Naive Bayes* memiliki sekitar 88%. Yadav & Pal (2012) mengeksplorasi nilai performa model algoritma klasifikasi *Decision Tree ID3, C4.5*, dan *CART* untuk memprediksi keberhasilan atau kelulusan nilai ujian akhir mahasiswa teknik. Model algoritma klasifikasi *Decision Tree C4.5* memiliki nilai akurasi yang paling tinggi yaitu 67,7778%.

Berdasarkan performa tiga model algoritma klasifikasi yang berbeda yaitu *Decision Tree, Naive Bayes*, dan *k-NN*, nilai performa bergantung pada jumlah data dan atribut (Tabel 1). Maka dari itu, evaluasi model algoritma klasifikasi menjadi sangat penting untuk mengetahui nilai performa yang dihasilkan pada suatu kasus. Misalnya pada kasus yang akan dievaluasi pada penelitian ini yang terdiri dari 49.735 *records* dan 47 atribut (Gambar 1). Pada penelitian ini juga, model algoritma klasifikasi *Decision Tree* yang terdapat banyak metode titik pembagi dahan (*Splitting Optimal Point*) diantaranya adalah *Gini Index(CART)*, *Information Gain (ID3)*, *Gain Ratio (C4.5)* (Yadav & Pal, 2012) akan dieksplorasi, dan untuk *k-NN* nilai *k* atau jarak *encludian* pada model algoritma tersebut mempengaruhi nilai performa.

**Tabel 1. Penelitian terkait**

Penelitian	Jumlah Atribut dan Record	Algoritma Terbaik
Durairaj & Deepika (2015)	7130 atribut dan 72 records	Naive Bayes
Shaikh, dkk. (2015)	14 Atribut dan 294 records	k-NN
Fitri (2014)	Ecoli (8 atribut dan 336 records), Yeast (9 atribut dan 1.484 records)	Naive Bayes & k-NN
Karim dan Rahman (2013)	17 atribut dan 4521 records	Decision Tree C4.5
Yadav dan Pal (2012)	17 atribut dan 90 records	Decision Tree C4.5
Ashari, dkk.(2013)	13 atribut dan 8.630 records	Naive Bayes

## 2.2. Decision Tree

Salah satu teknik klasifikasi yang paling populer digunakan adalah *decision tree* (Kamber & Han, 2006). *Decision tree* merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (*tree*) di mana setiap *node* merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas.

### 2.2.1. Information Gain

*Information Gain* merupakan suatu ukuran korelasi pada model parametrik yang menggambarkan ketergantungan antara dua peubah acak X dan Y. Metode *splitting optimal point* ini biasanya digunakan pada model algoritma ID3 (Gorunescu, 2011). *Information Gain* dapat dilihat pada persamaan (1), dimana persamaan (2) adalah *split info* yang didapat dari persamaan (3). *m* mempresentasikan banyaknya nilai yang berbeda atribut label kelas yang akan mendefinisikan kelas yang berbeda,  $C_i$  ( $i = 1, 2, \dots, m$ ).  $S_i$  mrepresentasikan jumlah sampel dalam himpunan sampel  $S$  (berisi  $s$  sampel) yang masuk kelas  $C_i$ .  $p_i$  merepresentasikan eluang bahwa suatu sampel akan masuk ke kelas  $C_i$  dan diestimasi dengan  $ssi$ .

$$\text{Gain}(A) = I(S_1, S_2, S_3, \dots, S_m) - E(A) \quad (1)$$

$$S_1, S_2, S_3, \dots, S_m \quad (2)$$

$$I(S_1, S_2, S_3, \dots, S_m) = -\sum_{i=1}^m p_i \text{Log}_2(p_i) \quad (3)$$

### 2.2.2. Gain Ratio

*Gain Ratio* merupakan modifikasi dari *Information Gain* untuk mengurangi bias atribut yang memiliki banyak cabang biasanya digunakan pada model algoritma *C4.5* (Gorunescu, 2011). *Gain Ratio* memiliki rumus seperti pada persamaan (4), dimana rumus *split info* seperti pada persamaan (4) dengan  $m$  menyatakan banyaknya *split*. Jenis *split* yang dipilih adalah *split* yang memiliki nilai *Gain Ratio* yang terbesar.

$$\text{Gain Ratio} = \frac{\text{Gain}}{\text{Split Info}} \quad (4)$$

### 2.2.3. Gini Index

Salah satu kriteria yang dapat digunakan untuk menentukan titik pemecah terbaik (*optimal splitting point*) adalah *Gini Index*, yang biasanya digunakan dalam algoritma CART (C&RT) dan SPRINT (Gorunescu, 2011). *Gini Index* atribut  $t$  untuk data dengan  $m$  kelas didefinisikan pada persamaan (5). Bila data dipecah terhadap  $A$  menjadi dua subset  $D1$  dan  $D2$ , maka *Gini Index* didefinisikan pada persamaan (6). Berdasarkan banyaknya *edge* keluar dari suatu atribut, maka terdapat dua jenis pemisahan yaitu *binary split* yang menghasilkan dua buah *edge* keluar dan *multyway split* yang menghasilkan lebih dari dua *edge* keluar.

$$\text{Gini}(t) = 1 - \sum_{i=1}^m p_i^2 \quad (5)$$

$$\text{Gini}_A(D) = \frac{|D1|}{|D|} \text{Gini}(D1) + \frac{|D2|}{|D|} \text{Gini}(D2) \quad (6)$$

### 2.3. Algoritma $k$ -NN

Algoritma  $k$ -NN adalah suatu metode yang menggunakan algoritma *supervised* (Kamber & Han, 2006). *Euclidean distance* adalah jarak yang paling umum digunakan pada data numerik yang didefinisikan pada persamaan (7) (Kamber & Han, 2006), dimana  $d(x_i, x_j)$  merepresentasikan *Euclidean distance*;  $(x_i)$  merupakan *record* ke- $i$ ;  $(x_j)$  adalah *record* ke- $j$ ;  $(a_r)$  adalah data ke- $r$ ; dan  $ij: 1, 2, 3, \dots, n$ .

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (7)$$

### 2.4. Naive Bayes

*Naive bayes* merupakan salah satu metode pada teknik klasifikasi dan termasuk dalam *classifier* statistik yang dapat memprediksi probabilitas keanggotaan *class*. *Naive Bayes* berprinsip pada teori *bayes*. NBC mengasumsikan bahwa nilai atribut pada sebuah *class* adalah independen terhadap nilai pada atribut yang lain (Kamber & Han, 2006). Klasifikasi *Bayesian* ini dihitung berdasarkan teorema *Bayes* (persamaan (8)). Berdasarkan persamaan (8) kejadian  $H$  merepresentasikan sebuah kelas dan  $X$  merepresentasikan sebuah atribut.  $P(H)$  disebut *prior probability H*, contoh dalam kasus ini adalah probabilitas kelas yang mendeklarasikan normal.  $P(X)$  merupakan *prior probability X*, contoh untuk probabilitas sebuah atribut *protocol\_type*.  $P(H|X)$  adalah *posterior probability* yang merefleksikan probabilitas munculnya kelas normal terhadap data atribut *protocol\_type*.  $P(X|H)$  menunjukkan kemungkinan munculnya prediktor  $X$  (*protocol\_type*) pada kelas normal. Dan begitu juga seterusnya untuk proses menghitung probabilitas keempat kelas lainnya.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (8)$$

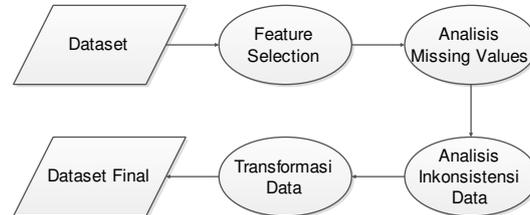
## 3. Dataset

Penelitian ini menggunakan *dataset* yang berasal dari *UCI Machine Learning* dan dari 55 atribut dan 101.776 *records*. Namun data-data tersebut masih terdapat data yang *redundant*, *noisy*, dan masih banyak *missing value*. Untuk keterangan tentang atribut dan deskripsi lengkap dapat dilihat pada penelitian (Strack, dkk., 2014). Tahap *preprocessing* perlu dilakukan karena data yang akan diadopsi dari *UCI Machine Learning* tersebut sebagai acuan perhitungan performa memiliki nilai atau data yang bersifat *redundant*, *noisy*, *multiple* (ganda), dan tidak komplit. Tahapan preparasi data dapat dilihat pada Gambar 1.

Gambar 1 menunjukkan tahapan proses *preprocessing* yang dilakukan untuk menghilangkan *noise* yang ada pada *dataset*. Tahap pertama yang dilakukan adalah *feature selection* atau pemilihan *feature* yang akan digunakan. Atribut *payer code* dan *patient code* dihapus karena tidak berhubungan dengan penyakit diabetes.

Kemudian tahap selanjutnya adalah analisis *missing values*. Pada *dataset* yang tersedia terdapat empat jumlah atribut yang *record* atau datanya mengandung *missing values*. Atribut-

atribut tersebut adalah *weight* 97%, *race* 2%, *diagnosis* 3 1%, dan *medical specialty* 47%. Fitur *weight* memiliki data *missing values* yang sangat besar sehingga tidak dimungkinkan untuk diganti dengan data-data yang dapat dilakukan dengan teknik-teknik statistik sehingga atribut ini dipilih oleh penulis merupakan kandidat atribut yang dihapus.



**Gambar 1. Tahap Preparasi Data**

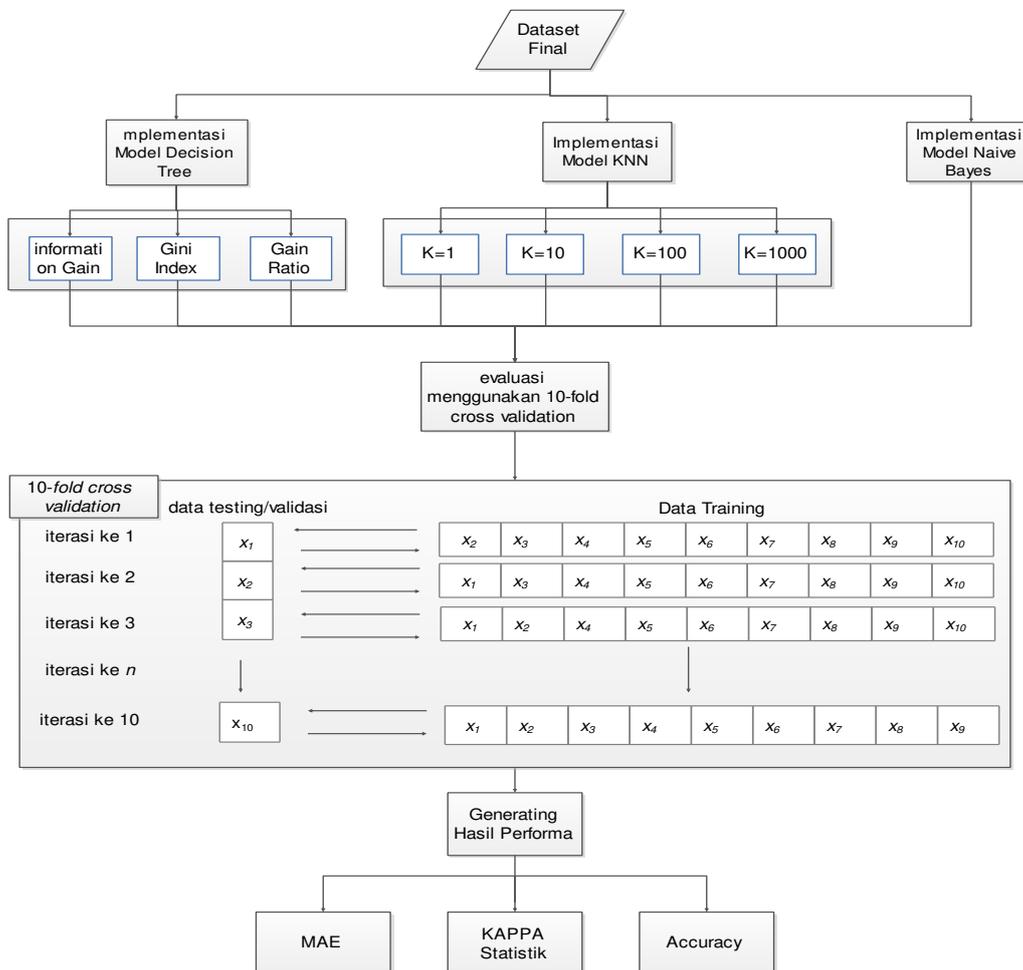
Namun dalam proses analisis duplikat data atau yang diwakili tahap proses analisis inkonsisten data, tidak ditemukannya data yang duplikat setelah dilakukan analisis terhadap *id encounter* sebagai dasar penentuan *id label*. Kemudian pada tahap transformasi data, *values* yang terdapat pada *diagnosis 1*, *diagnosis 2*, dan *diagnosis 3* akan diganti dengan klasifikasi kode diagnosis sesuai dengan kode *icd9 (International Statistical Classification of Diseases and Related Health Problems)* yang merupakan standar klasifikasi statistik penyakit internasional yang berhubungan dengan masalah penyakit (Strack, dkk., 2014). Setelah dilakukan tahap-tahap *preprocessing* tersebut, *dataset* final yang dijadikan data untuk skenario percobaan menjadi 47 atribut dan 49.735 *records*. Deskripsi lengkap tentang atribut dapat dilihat pada Tabel 2.

**Tabel 2. Daftar Atribut dan Deskripsinya**

No	Nama Atribut	Tipe	Deskripsi
1	<i>Encounter ID</i>	Polynominal (ID)	Nomor Kunjungan sebagai ID
2	<i>Race</i>	Polynominal	<i>Value</i> : <i>Caucasian, Asian, African American, Hispanic, dan Other</i>
3	<i>Gender</i>	Polynominal	Nilai : <i>Male, Female, dan Unknown</i>
4	<i>Age</i>	Polynominal	<i>Value</i> : Dikelompokan dengan interval 10 tahun (0-10, 10-20,...,90-100)
5	<i>Admission Type</i>	Polynominal	<i>Value</i> : <i>Emergency, urgent, elective, newborn, dan not available.</i>
6	<i>Discharge Disposition</i>	Polynominal	<i>Value</i> : <i>discharge to home, expired, not available.</i>
7	<i>Admission Source</i>	Polynominal	<i>Value</i> : <i>physician referral, emergency room, dan transfer from hospital.</i>
8	<i>Time in Hospital</i>	Polynominal	Jumlah atau lama pasien dirawat di rumah sakit dari pendaftaran sampai pelepasan keluar dari rumah sakit.
9	<i>Medical Specialty</i>	Polynominal	Penanganan khusus misalnya <i>cardiology, internal medicine, dan dll.</i>
10	<i>Number of lab procedures</i>	Polynominal	Jumlah tes-tes lab yang dilakukan pada sekali kunjungan
11	<i>Number of procedures</i>	Polynominal	Jumlah prosedur pada satu kali kunjungan
12	<i>Number of medications</i>	Polynominal	Jumlah obat-obatan untuk pasien yang diberikan dalam sekali kunjungan
13	<i>Number of outpatients visits</i>	Polynominal	Jumlah Kunjungan pasien rawat jalan dalam sekali proses pengobatan
14	<i>Number of emergency visits</i>	Polynominal	Jumlah kunjungan emergency ketika sedang dalam tahap pengobatan
15	<i>Number of inpatients visits</i>	Polynominal	Jumlah kunjungan pasien rawat inap yang sedang dalam tahap pengobatan
16	<i>Diagnosis 1</i>	Polynominal	Diagnosis utama; terdapat 848 nilai berbeda.
17	<i>Diagnosis 2</i>	Polynominal	Diagnosis kedua; terdapat 923 nilai berbeda
18	<i>Diagnosis 3</i>	Polynominal	Diagnosis tambahan yang mendukung diagnosis kedua; terdapat sebanyak 954 nilai yang berbeda
19	<i>Number of diagnoses</i>	Polynominal	Jumlah diagnosa yang di-inputkan ke dalam sistem
20	<i>Glucose serum test result</i>	Polynominal	Mengindikasikan rentang hasil; nilai, >200,>300, normal, dan none (tidak ada).
21	<i>A1c test result</i>	Polynominal	Lebih dari normal diindikasikan dengan nilai >8%, jika dalam batas normal yaitu antara 7-8% dikodekan dengan normal, sedangkan jika kekurangan diindikasikan dengan kode <7%, dan <i>none</i> jika tidak dihitung.
22	<i>Change of medication</i>	Polynominal	Mengindikasikan jika ada perubahan dalam pengobatan (baik dosis obat atau obat yang dipakai), nilai; No (jika tidak ada perubahan) atau Change (jika berubah)
23	<i>Diabetes Medication</i>	Polynominal	Mengindikasi jika ada pengobatan diabetes lain yang diresepkan. Nilai : Yes dan No
24	<i>24 features for medications</i>	Polynominal	Nama-nama obat seperti metformin, repaglide, nateglinide, chlorpropamide, dan lain-lain.
25	<i>Status</i>	Polynominal (Class Label)	Nilai >30 untuk pasien yang kembali dirawat inap lebih dari 30 hari, Nilai >30 untuk pasien yang kembali dirawat inap kurang dari 30 hari, dan No untuk yang tidak direkamisi.

#### 4. Metode Penelitian dan Skenario Percobaan

Metode penelitian yang akan dilakukan dengan adopsi dari pada tahapan proses KDD (*Knowledge Discovery in Data mining*). Tahapan KDD antara lain (1) *Data Selection* yaitu memilih data yang dapat digunakan untuk penelitian. (2) *Preprocessing* Data yaitu membersihkan data yang tidak lengkap atau mengandung duplikasi dan memeriksa data yang tidak konsisten. (3) Transformasi data yaitu data diubah menjadi bentuk yang sesuai untuk penambangan data. (4) Aplikasi teknik penambangan data yaitu proses inti dimana teknik penambangan data diterapkan untuk mengekstrak pola-pola tertentu pada data. (5) Evaluasi pola yang ditemukan yaitu hasil dari teknik penambangan data berupa pola-pola yang khas maupun prediksi evaluasi untuk menilai hipotesa yang ada. (6) Presentasi pengetahuan yaitu menggunakan teknik visualisasi untuk menampilkan hasil penambangan data kepada pengguna (Sari, dkk., 2016). Untuk tahap (1), (2), dan (3) dijelaskan pada Bab 3, sedangkan untuk tahap (4), dan (5) akan dijelaskan pada Bab 4 dan Untuk tahap (6) Akan dijelaskan pada Bab 5. Gambar 2 merupakan skenario percobaan yang dilakukan pada penelitian ini. Proses skenario ini menjelaskan tahapan KDD tahap (4), (5), dan (6).



**Gambar 2. Skenario Percobaan**

Metode penelitian yang akan dilakukan adalah metode *supervised learning* yaitu pembelajaran terarah berdasarkan data historis terkait readmisi pasien diabetes. Skenario percobaan dimulai dari menginputkan *dataset* final yang diambil setelah dilakukannya tahap *preprocessing* (Gambar 2). Kemudian *dataset* tersebut di-*import* ke *machine learning* Rapidminer 7.0.0. Setelah itu, *dataset* diuji coba dengan mengimplementasikan model-model algoritma klasifikasi misalnya untuk uji coba pertama implementasi *Decision Tree* dengan

*parameter setting information gain*. Kemudian model algoritma tersebut divalidasi menggunakan teknik *10-fold cross validation* menggunakan *tool* pembelajaran mesin. Teknik validasi ini dimulai dari membagi subset *dataset* menjadi 10 subset ( $x_1, x_2, \dots, x_{10}$ ). Kemudian ketika  $x_1$  menjadi data validasi atau *data testing* dan  $x_2, x_3, \dots, x_{10}$  akan menjadi *data training* kemudian data tersebut diiterasi secara berurutan sampai dengan *data testing*  $x_{10}$  dengan *data training*  $x_1, x_2, \dots, x_9$ . Setelah divalidasi kemudian *machine learning* Rapidminer akan menghasilkan *output* berupa performa algoritma dengan data aktual yang dipaparkan melalui hasil pengukuran setiap algoritma.

Perlakuan serupa juga diberikan untuk model-model algoritma yang lainnya. Setiap algoritma diberikan *parameter setting* atau metode yang berbeda-beda sesuai dengan kapabilitas masing-masing algoritma. Misalnya untuk *Decision Tree*, penerapan metode atau teknik *Entropy* atau *Splitting Optimal Point* yang diterapkan adalah berdasarkan *Information Gain*, *Gain Ratio*, dan *Gini Index* yang merupakan adopsi dari penelitian Yadav & Pal (2012). Sedangkan untuk model algoritma klasifikasi *k-NN* parameter  $k$  yang digunakan adalah  $k=1, 10, 100, 1000$  karena performa algoritma *k-NN* sangat bergantung pada nilai  $k$  dan *dataset* yang digunakan untuk validasi memiliki *record* yang besar sehingga nilai  $k$  yang digunakan menyesuaikan dengan jumlah data. Dan kemudian pengujian yang terakhir adalah mengimplementasikan model algoritma *Naive Bayes*.

Pemilihan model algoritma klasifikasi ini berdasarkan penelitian Durairaj & Deepika (2015) yang menyatakan bahwa algoritma *Naive Bayes* memiliki kelebihan sangat akurat dalam mengklasifikasikan data dengan jumlah atribut yang banyak. Sedangkan penelitian Karim & Rahman (2013) menunjukkan bahwa algoritma *Decision Tree* memiliki nilai *error* atau *error rate* yang kecil. Kemudian Shaikh, dkk. (2015) dalam hasil penelitian menunjukkan bahwa model algoritma klasifikasi *k-NN* memiliki nilai performa akurasi yang baik jika diterapkan pada jumlah *record* yang besar. Maka dari itu dalam penelitian ini, tiga algoritma ini akan dievaluasi dan divalidasi guna pencarian atau penentuan terkait readmisi pasien diabetes.

## 5. Hasil Percobaan

Setelah dilakukannya skenario percobaan dengan mengimplmentasikan seluruh algoritma klasifikasi dengan berbagai *parameter setting* menggunakan *machine learning tool* Rapidminer 7.0.0, maka didapatkan hasil yang dipaparkan pada Tabel 3.

**Tabel 3. Hasil performa evaluasi model algoritma**

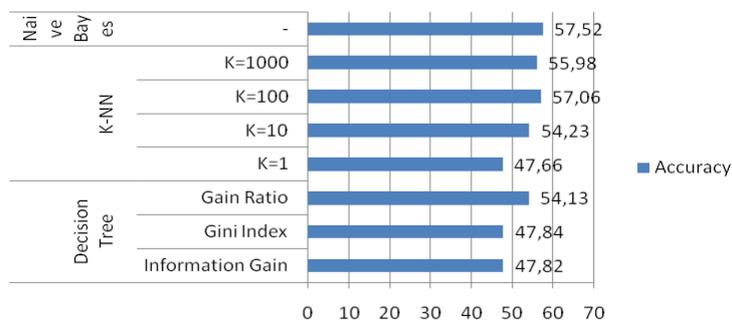
Model Algoritma Klasifikasi	Parameter Setting	Accuracy	MAE	Kappa Statistik
<i>Decision Tree</i>	<i>Information Gain</i>	47,82	0,525	0,087
	<i>Gini Index</i>	47,84	0,525	0,086
	<i>Gain Ratio</i>	54,13	0,524	0,138
<i>K-NN</i>	$k=1$	47,66	0,523	0,072
	$k=10$	54,23	0,528	0,109
	$k=100$	57,06	0,534	0,087
	$k=1000$	55,98	0,543	0,032
<i>Naive Bayes</i>	-	57,52	0,512	0,182

### 5.1. Accuracy

*Accuracy* dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Kamber & Han, 2006). Rumus *Accuracy* dapat didefinisikan pada persamaan (9), dimana TP merupakan jumlah *record/tuple* positif yang dilabelkan secara benar oleh model algoritma klasifikasi, TN merupakan jumlah *record/tuple* negatif yang dilabelkan secara salah oleh model algoritma klasifikasi, dan P merupakan total semua *record* yang dievaluasi. Semakin tinggi level *accuracy*-nya, maka dapat dikatakan semakin efektif model algoritma klasifikasi tersebut (Mittal & Gill, 2014).

$$Accuracy = \frac{TP+TN}{P} \quad (9)$$

Gambar 3. menunjukkan bahwa model algoritma klasifikasi terakurat pada percobaan terhadap pada percobaan ini adalah model algoritma klasifikasi *Naive Bayes* dengan jumlah akurasi sebesar 57,52%. Sedangkan yang paling tidak akurat didapat dari model algoritma *k-NN* dengan implementasi nilai  $k=1$  dengan nilai akurasi sebesar 47,66%. Penelitian Durairaj & Deepika (2015) menyebutkan bahwa algoritma klasifikasi *Naive Bayes* memiliki nilai performa akurasi yang baik dibandingkan *k-NN* dan *Decision Tree* setelah mengevaluasi *dataset* yang terdiri dari 7.130 atribut dan 72 records dan penelitian Fitri (2014) yang menyatakan bahwa model algoritma *Naive Bayes* juga memiliki performa akurasi tertinggi pada *dataset* *Ecoli* (8 atribut dan 336 records) dan *Yeast* (9 atribut dan 1.484 records). Sehingga berdasarkan hasil temuan yang ditunjukkan pada Gambar 3, selain memiliki performa akurasi yang baik ketika dievaluasi dengan jumlah atribut yang banyak, model algoritma klasifikasi *Naive Bayes* juga memiliki tingkat keakuratan yang baik jika digunakan untuk jumlah record yang besar. Misalnya penelitian yang dilakukan ini, performa akurasi algoritma *Naive Bayes* memiliki nilai performa akurasi yang paling baik dibandingkan dengan *k-NN* dan *Decision Tree* ketika dievaluasi dengan *dataset* yang memiliki jumlah record 49.735 dan 47 atribut.



Gambar 3. Grafik Hasil Percobaan nilai performa akurasi

**5.2. Mean Absolute Error (MAE)**

*Mean Absolute Error* (MAE) adalah sebuah kuantitas yang digunakan untuk menghitung seberapa dekat atau akurat prediksi terhadap hasil yang ditemukan. MAE menggambarkan atau menunjukkan kapabilitas dari sebuah algoritma (Mittal & Gill, 2014). Semakin kecil nilai MAE, maka semakin kuat kapabilitas algoritma tersebut.

Gambar 4 merupakan hasil performa MAE yang mewakili kekuatan atau *robustness* dari berbagai macam model algoritma klasifikasi yang dipilih terhadap 49.735 records dan 47 atribut. Dari Gambar 4 dapat diketahui bahwa model algoritma *Naive Bayes* memiliki nilai performa MAE yang paling kecil dibandingkan dengan dengan model-model algoritma klasifikasi yang lainnya yaitu 0,512. Sedangkan model algoritma klasifikasi dengan nilai performa MAE terbesar adalah model *k-NN* dengan implementasi nilai  $k=1000$  yaitu sebesar 0,543. Ini berarti model algoritma klasifikasi *Naive Bayes* merupakan model algoritma klasifikasi yang kapabilitas khususnya penanganan *error* paling baik dan kapabilitas terburuk didapatkan oleh model algoritma klasifikasi *k-NN* dengan  $k=1000$ . Hal ini ditunjukkan pada Gambar 3 bahwa model algoritma *Naive Bayes* memiliki nilai akurasi tertinggi.

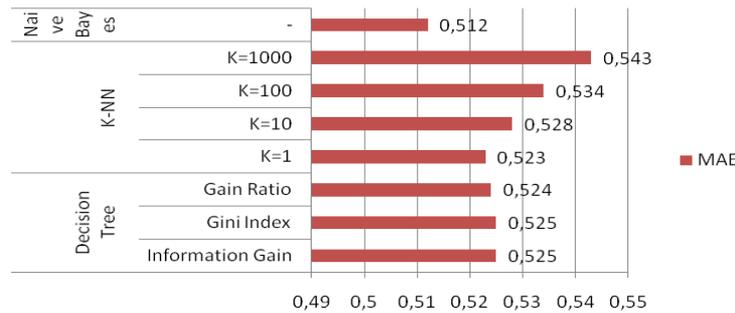
Namun, jika dilihat di dalam penelitian Fitri (2015), algoritma *k-NN* memiliki nilai MAE terbaik ketika dievaluasi pada *dataset* *Yeast* yang terdiri 9 atribut dan 1.484 records namun pada *dataset* *Ecoli* (8 atribut dan 336 records) *Naive Bayes* memiliki nilai performa yang lebih baik. Berdasarkan hasil penelitian tersebut dan yang ditunjukkan pada Gambar 4 menunjukkan bahwa model algoritma *k-NN* tidak selalu memiliki nilai performa MAE yang baik. Hal ini mengarahkan pada perlunya evaluasi *dataset* agar pemilihan model algoritma klasifikasi yang akan diimplementasikan pada suatu kasus dapat optimal.

**5.3. Kappa Statistic**

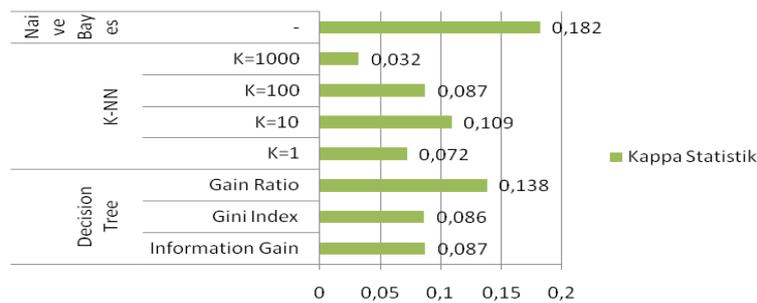
*Kappa Statistic* adalah analisis statistikal yang didasarkan pada kecocokan penafsiran

atau derajat kesepakatan (*degree of agreement*) untuk data kualitatif. Pada dasarnya *kappa statistic* menunjukkan analisis diantara kelas-kelas yang berbeda. Semakin tinggi *kappa statistic*, maka akan dipertimbangkan sebagai performa yang mempunyai kriteria atau kinerja yang bagus (Mittal & Gill, 2014).

Gambar 5 adalah hasil pengujian performa model algoritma klasifikasi berdasarkan nilai *kappa statistic*. Dari hasil pengujian tersebut nilai *kappa* tertinggi dihasilkan dari model algoritma *Naive Bayes* dengan nilai *kappa statistic* sebesar 0,182. Sedangkan nilai terendah dihasilkan dari model algoritma *k-NN* dengan nilai  $k=1000$  yaitu 0,032. Sehingga kriteria kapabilitas terbaik secara statistik dimiliki oleh model algoritma *Naive Bayes*.



Gambar 4. Grafik Hasil Percobaan nilai performa MAE



Gambar 5. Grafik Hasil Percobaan nilai performa *kappa statistic*

## 6. Kesimpulan dan Saran

Berdasarkan hasil dari pengujian-pengujian dari implementasi model-model klasifikasi *Decision Tree*, *k-NN*, dan *Naive Bayes* dengan berbagai *parameter setting* dengan *dataset* evaluasi terkait readmisi pasien diabetes yang terdiri dari 47 atribut dan 49.735 *records* didapatkan hasil bahwa model algoritma klasifikasi terbaik diperoleh dari hasil uji coba implementasi model algoritma klasifikasi *Naive Bayes*. Hal ini dibuktikan dengan data hasil pengujian dari model algoritma klasifikasi *Naive Bayes* yaitu dengan nilai akurasi sebesar 57,52%, MAE sebesar 0,512, dan *kappa statistic* sebesar 0,182. Hal ini membuktikan bahwa model *Naive Bayes* memiliki tingkat keakurasian, *Error Rate*, dan kapabilitas secara statistik paling baik dari model-model algoritma klasifikasi yang lainnya seperti *k-NN* dan *Decision Tree*. Sehingga rekomendasi untuk perhitungan kualitas medikasi terkait pasien diabetes disarankan untuk mengimplementasi model algoritma klasifikasi *Naive Bayes*.

Penelitian ini masih memiliki keterbatasan-keterbatasan. Maka dari itu, rencana penelitian yang akan dilanjutkan pada penelitian berikutnya adalah mengevaluasi jumlah *record* yang terhadap performa model algoritma klasifikasi. Kemudian implementasi model-model dengan pendekatan statistik dan sekuensial akan diteliti lebih lanjut.

## Referensi

- American Diabetes Association. (2013). Economic Costs of Diabetes in the U.S. in 2012. *Diabetes Care*, 36(4), pp. 133-146.
- Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance Comparison between Naïve Bayes,

- Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 4, No. 11, pp. 33-29.
- Dungan, K. (2012). The effect of diabetes on hospital readmissions. *Jurnal of Diabetes & Science Technology*. Volume 6, No. 5., pp. 1045-1052.
- Durairaj, M., & Deepika, R. (2015). Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer. *International Journal of Advanced Research in Computer Science and Software Engineering*; Volume 5, No. 8, pp. 787-791.
- Fitri, S., (2014). Perbandingan Kinerja Algoritma Klasifikasi Naïve Bayesian, Lazy-Ibk, Zero-R, Dan Decision Tree- J48. *Jurnal Dasi Vol. 15 No. 1*. pp. 33-37.
- Gorunescu, F. (2011). *Data mining Concept Model and Techniques*. Berlin: Springer.
- Iskandar, D. (2014). *Analisis Kejadian Rawat Inap Ulang (Readmission) di RSUD Prof. Dr. Margono Soekarjo Purwokerto*. Purwokerto: Universitas Jendral Sudirman.
- Kamber, M., & Han, J. (2006). *Data mining; Concepts and Techniques Second Edition*. San Francisco: Morgan Kaufmann Publishers.
- Karim, M., & Rahman, R. M. (2013). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *Journal of Software Engineering and Applications Volume. 6*, pp. 196-206.
- Kaur, B. & Singh, W. (2014). Review on Heart Disease Prediction System using *Data mining Techniques*. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC) Volume: 2 Issue: 10*. pp. 3003-3008.
- Mittal, P., & Gill, N. S. (2014). A Comparative Analysis Of Classification Techniques On Medical Data Sets. *IJRET: International Journal of Research in Engineering and Technology, Volume: 03 Number: 06*, pp. 454-460.
- Rahman, R. M., & Afroz, F. (2013). Comparison of Various Classification Techniques Using Different *Data mining Tools for Diabetes Diagnosis*. *Journal of Software Engineering and Applications Vol.6*, pp.85-97.
- Ramirez, S.P., McCullough, K.P., Thumma J.R., Nelson, R.G., Morgenstern, H., Gillespie B.W., Inaba, M., Jacobson S.H., Vanholder, R., Pisoni, R.L., Port F.K., Robinson B.M., (2012). Hemoglobin A1c Levels and Mortality in the Diabetic Hemodialysis Population Findings from the Dialysis Outcomes and Practice Patterns Study (DOPPS). *Diabetes care, 35(12)*, pp. 2527-2532.
- Rubin, D. J. (2015). Hospital readmission of patients with diabetes. *Current Diabetes Reports. 15(4)*, pp. 1-9.
- Sari, M.K., Ernawati, & Wisnubhadra, I. (2016). Pembangunan Aplikasi Klasifikasi Mahasiswa Baru untuk Prediksi Hasil Studi Menggunakan Naïve Bayes Classifies. *Jurnal Buana Informatika Volume 7 Nomor 2*. pp. 135-142.
- Shaikh, A., Mahoto, N., Khuhawar, F., & Memon, M. (2015). Performance Evaluation of Classification Methods for Heart Disease Dataset. *Sindh University Research Journal-SURJ (Science Series). 47(3)*, pp. 389-394.
- Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J., Clore, J.N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international Hindawi*, pp.1-11.
- Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications, 36*, pp. 8610–8615.
- Tomar, D., & Agarwal, S. (2013). A Survey on *Data mining Approaches for Healthcare*. *International Journal of Bio-Science and Bio-Technology Vol.5, No.5.*, pp. 241-266.
- Yadav, S. K., & Pal, S. 2012. *Data mining: A Prediction for Performance Improvement of Engineering Students using Classification*. *World of Computer Science and Information Technology Journal (WCSIT) Vol. 2, No. 2*, pp. 51-56.
- Yoon, H., Park, C.S., Kim, J. S., & Baek, J.G. (2013). Algorithm learning based neural network integrating feature selection and classification. *Expert Systems with Applications, 40(1)*, pp. 231-241.