

Terbit online pada laman : <http://teknosi.fti.unand.ac.id/>

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |



Artikel Penelitian

Prediksi Diabetes Menggunakan Algoritma Naïve Bayes dan Greedy Forward Selection

Fitriyani

Universitas ARS, Jl. Sekolah Internasional 1-2, Bandung, 40282, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel:

Diterima Redaksi: 08 Agustus 2021

Revisi Akhir: 28 Agustus 2021

Diterbitkan Online: 31 Agustus 2021

KATA KUNCI

Prediksi Diabetes,
Naïve Bayes,
Seleksi Fitur,
Greedy Forward Selection

KORESPONDENSI

E-mail: fitriyani@ars.ac.id

ABSTRACT

Diabetes merupakan penyakit yang mengancam kehidupan dengan pertumbuhan tercepat dan jika tidak diobati atau diidentifikasi akan menyebabkan komplikasi lain. Diabetes setiap tahunnya mengakibatkan kematian sebanyak 3.8 juta jiwa dan telah mempengaruhi 422 juta orang di seluruh Dunia. Dataset diabetes yang digunakan dalam penelitian ini adalah dataset publik, dataset ini akan diolah menggunakan model yang diusulkan. Dataset diabetes memiliki permasalahan seperti adanya fitur-fitur yang tidak relevan, fitur-fitur yang tidak relevan dapat menurunkan kinerja dari model yang digunakan. Seleksi fitur Greedy Forward Selection adalah seleksi fitur yang sangat efisien dan cepat dalam prosesnya. Algoritma Naïve Bayes merupakan algoritma yang mudah dan sederhana ketika diimplementasikan. Hasil penelitian menunjukkan bahwa model Naïve Bayes dan Greedy Forward Selection mendapatkan nilai akurasi tertinggi sebesar 91.73%, sedangkan model Naïve Bayes tanpa seleksi fitur Greedy Forward Selection hanya mendapat nilai akurasi sebesar 87.69%.

1. PENDAHULUAN

Diabetes adalah salah satu penyakit kronis yang mengancam kehidupan dengan pertumbuhan tercepat yang telah mempengaruhi 422 juta orang di seluruh dunia menurut laporan dari Organisasi Kesehatan Dunia (WHO) pada tahun 2018 [1]. Di Indonesia menurut Riset Kesehatan Dasar (Riskesdas) bahwa kenaikan Diabetes sebesar 6.9% dari total populasi orang diatas 15 tahun [2]. Penyakit diabetes dapat disebut juga penyakit paling mematikan dan kronis yang menyebabkan peningkatan gula darah (glukosa). Glukosa merupakan sumber energi utama bagi sel tubuh manusia. Glukosa yang menumpuk di dalam darah akibat tidak diserap sel tubuh dengan baik dapat menimbulkan berbagai gangguan organ tubuh. Jika diabetes tidak diobati dan diidentifikasi, banyak komplikasi yang akan terjadi [3]. Diabetes menyebabkan penyakit atau komplikasi lain yang setiap tahunnya mengakibatkan kematian 3,8 juta jiwa. Komplikasi lain yang lebih sering terjadi dan mematikan akibat diabetes adalah serangan jantung dan stroke. Sebagian besar kematian terjadi

karena kenaikan kadar glukosa secara terus menerus sehingga mengakibatkan rusaknya pembuluh darah, saraf dan struktur internal lainnya [4]. Diabetes terbagi menjadi 2 (dua), yaitu diabetes tipe 1 dan tipe 2. Diabetes tipe 1 terjadi ketika sistem kekebalan tubuh secara keliru menyerang sel di pankreas dan sangat sedikit insulin yang dikeluarkan oleh tubuh atau terkadang bahkan tidak ada insulin yang dikeluarkan oleh tubuh. Sebaliknya, diabetes tipe 2 terjadi ketika tubuh kita tidak memproduksi insulin yang tepat atau tubuh menjadi resisten terhadap insulin [1]. Identifikasi awal merupakan satu-satunya cara untuk menghindari komplikasi [3] dan dapat mengobati diabetes secara cepat dan tepat.

Dataset yang digunakan dalam penelitian ini adalah dataset diabetes yang merupakan dataset publik dari UCI Repository. Penelitian menggunakan dataset publik dianjurkan karena sebanyak 64.79% penelitian di dunia menggunakan dataset publik dan sisanya 35.21% menggunakan dataset privat [5]. Penggunaan dataset publik dapat membuat penelitian berulang terbantahkan dan diverifikasi [6].

Algoritma Naïve Bayes adalah pengklasifikasi probabilitas sederhana yang menghitung sekumpulan probabilitas dengan menghitung frekuensi dan kombinasi nilai dalam kumpulan data yang diberikan [7]. Menurut [8] Klasifikasi menggunakan Naïve Bayes dapat menghasilkan akurasi yang tinggi dan cepat ketika diaplikasikan pada data yang besar serta mudah dalam penerapannya [9]. Seleksi fitur menggunakan Greedy Forward Selection sangat efisien, sederhana dan tidak membutuhkan waktu lama dalam prosesnya [10].

Banyak penelitian yang dilakukan sebelumnya dengan menggunakan algoritma yang sama, seperti pada penelitian [11] menggunakan algoritma K-NN, Naïve Bayes dan Decision Tree yang diimplementasikan pada 3 dataset yaitu Waether Nominal, Segment Challenge dan Supermarket. Hasil penelitian menunjukkan akurasi tertinggi pada algoritma KNN dan Decision Tree sebesar 100%. Penelitian terkait selanjutnya menggunakan dataset diabetes yaitu Pima Indians Diabetes Database (PIDD) yang diolah dengan algoritma Support Vector Machine, Naïve Bayes dan Decision Tree. Hasil dalam penelitian ini untuk akurasi tertinggi pada algoritma Naïve Bayes sebesar 76.30% [3]. Kemudian pada penelitian terkait lainnya [12] yang menggunakan dataset diabetes yaitu dataset Diabetes Mellitus, akan tetapi diolah dengan algoritma Random Forest, C4.5, REPTree dan Logistic Model Tree. Akurasi tertinggi ada pada algoritma Logistic Model Tree sebesar 79.31%. Selanjutnya penelitian [10] menggunakan algoritma Decision Tree dan Greedy Forward Selection untuk mengolah dataset cryotherapy dan immunotherapy. Hasil penelitian menunjukkan akurasi terbaik pada algoritma Decision Tree dengan Greedy Forward Selection sebesar 92.22%. Penelitian [13] menggunakan dataset Diabetes Mellitus yang diolah dengan algoritma Support Vector Machine, Artificial Neural Network, Decision Tree dan Naïve Bayes. Penelitian ini menghasilkan akurasi terbaik pada algoritma Support Vector Machine sebesar 82%. Pada penelitian [7] algoritma yang digunakan dalam penelitiannya adalah algoritma Artificial Neural Network dan Naïve Bayes yang diimplementasikan pada dataset Breast Cancer Coimbra. Hasil akurasi terbaik ada di algoritma Artificial Neural Network sebesar 86.95%. Pada penelitian ini diimplementasikan algoritma Naïve Bayes dan seleksi fitur Greedy Forward Selection untuk memprediksi pasien yang terkena diabetes dan melihat hasil kinerja dari algoritma yang diimplementasikan.

2. METODE

2.1. Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset diabetes yang diunduh pada situs UCI Repository. Dataset ini ada 520 record dan 17 atribut, dataset ini didapat dari hasil kuisioner pada pasien di Rumah Sakit Sylhet Bangladesh. Deskripsi atribut dataset diabetes dapat dilihat pada table 1.

Tabel 1. Deskripsi Atribut Diabetes

ATRIBUT	DESKRIPSI
Age	Jenis Kelamin
Gender	Usia
Polyuria	Sering buang air kecil
Polydipsia	Rasa haus yang berlebih
Sudden Weight Loss	Penurunan Berat Badan yang signifikan
Weakness	Kelelahan atau lemas
Polyphagia	Nafsu makan meningkat
Genital Thrush	Infeksi jamur pada Kelamin
Visual Blurring	Pandangan kabur atau tidak jelas
Itching	Gatal di bagian tubuh tertentu dan sulit sembuh
Irritability	Emosi tidak stabil
Delayed Healing	Luka sulit sembuh
Partial Paresis	Lumpuh atau merasa lemas
Muscle Stiffness	Badan tidak seimbang dan merasa kaku
Alopecia	Kebotakan atau rambut rontok
Obesity	Obesitas
Class	Negatif atau Positif

2.2. Naïve Bayes

Pengklasifikasi Naive Bayes adalah pengklasifikasi probabilistik sederhana berdasarkan penerapan teorema Bayes (dari statistik Bayesian) dengan asumsi penentuan nasib sendiri yang kuat (naif) [14]. Naïve Bayes dapat disebut juga dengan Simple Bayes dan Independence Bayes. Algoritma ini dapat memprediksi probabilitas keanggotaan kelas, seperti probabilitas pada data yang diberikan label kelas tertentu. Pengklasifikasi Naive Bayes menganggap bahwa ada (atau tidak adanya) fitur (atribut) tertentu dari suatu kelas tidak terkait dengan ada (atau tidak adanya) fitur lain ketika variabel kelas diberikan [11]. Berikut persamaan Naïve Bayes [14]:

$$p(C|F_1 \dots F_n) = \frac{p(C)p(F_1 \dots F_n|C)}{p(F_1 \dots F_n)} \quad (1)$$

Keuntungan menggunakan algoritma Naïve Bayes adalah:

1. Cepat dan model sangat terukur
2. Menyeimbangkan linear dengan jumlah prediktor dan baris
3. Prosedur Naïve Bayes adalah parallel
4. Naïve Bayes dapat digunakan untuk klasifikasi biner dan multiclass.

2.3. Seleksi Fitur

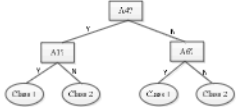
Seleksi fitur dapat mengurangi dimensional pada data dan untuk meningkatkan kinerja dari mesin pembelajaran. *Subset selection* merupakan metode *feature selection* (seleksi fitur), *subset selection* adalah menemukan *subset* terbaik (fitur), *subset* yang terbaik mempunyai jumlah dimensi yang paling berkontribusi pada akurasi [10]. Seleksi fitur juga dapat mengurangi dimensional pada data dan meningkatkan kinerja dari mesin pembelajaran sehingga dapat menyelesaikan permasalahan pada regresi dan klasifikasi [15].

2.4. Greedy Forward Selection

Subset selection merupakan metode feature selection (seleksi fitur), subset selection adalah menemukan subset terbaik (fitur), subset yang terbaik mempunyai jumlah dimensi yang paling berkontribusi pada akurasi [16]. Greedy Forward Selection

merupakan salah satu pendekatan dalam algoritma greedy pada metode seleksi atribut atau seleksi fitur [10]. Metode seleksi ini terdapat 2 yaitu Forward Selection dan Backward Elimination. Pada Tabel 2 dapat dilihat metode heuristic Greedy atau dapat juga disebut dengan metode pendekatan Greedy.

Tabel 2. Metode Heuristic Greedy [8]

Forward Selection	Backward Elimination	Decision Tree Induction
Atribut asli: {A1, A2, A3, A4, A5, A6} Pengurangan atribut asli: {} $\Rightarrow \{A1\}$ $\Rightarrow \{A1, A4\}$ \Rightarrow Pengurangan atribut: {A1, A4, A6}	Atribut asli: {A1, A2, A3, A4, A5, A6} $\Rightarrow \{A1, A3, A4, A5, A6\}$ $\Rightarrow \{A1, A4, A5, A6\}$ \Rightarrow Pengurangan Atribut: {A1, A4, A6}	Atribut asli: {A1, A2, A3, A4, A5, A6}  \Rightarrow Pengurangan Atribut: {A1, A3, A6}

2.5. K-Fold Cross Validation

Dalam *k-fold cross validation*, dataset dibagi secara acak sebanyak K bagian, $\chi_i = i, \dots, K$. Untuk menghasilkan masing-masing bagian, salah satu bagian K sebagai validasi dan menggabungkan yang tersisa dari bagian $K-1$ menjadi data *training* [17]. Berikut persamaan dari *k-fold cross validation*:

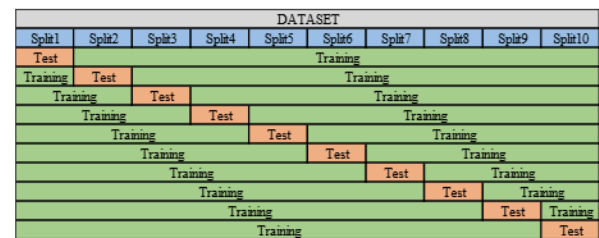
$$v_1 = x_1 T_1 = x_2 \cup x_3 \cup \dots \cup x_k \quad (2)$$

$$v_2 = x_2 T_2 = x_1 \cup x_3 \cup \dots \cup x_k$$

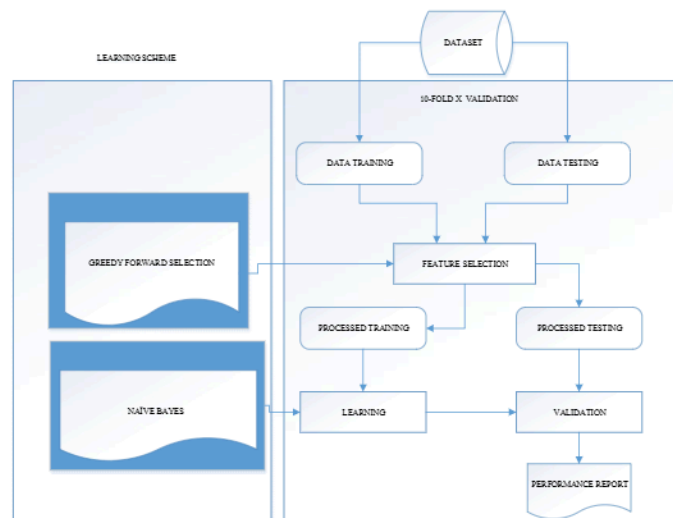
...

$$v_k = x_k T_k = x_1 \cup x_2 \cup \dots \cup x_{k-1}$$

Penggunaan 10 *fold cross validation* karena metode ini merupakan metode standar dalam praktek [18]. Dataset dibagi menjadi 10 bagian, dimana dataset ini terdiri dari 2 jenis, yaitu data *training* dan data *testing*. Gambar 1 merupakan ilustrasi dari 10 *fold cross validation*.



Gambar 1. Model 10 Fold Cross Validation



Gambar 2. Kerangka Penelitian

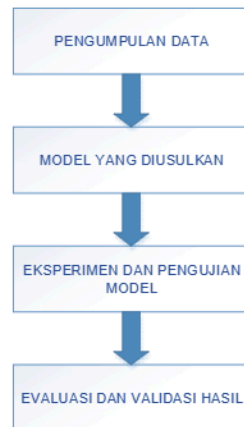
2.6. Kerangka Penelitian

Pada penelitian ini dataset diabetes akan dibagi menjadi 10 bagian menggunakan *k-fold cross validation*, dimana bagian ini terdiri dari data *training* dan data *testing*. Data *training* dan data *testing* akan diolah menggunakan seleksi fitur Greedy Forward

Selection, seleksi fitur ini untuk menemukan fitur yang paling relevan, kemudian hasil dari proses seleksi ini selanjutnya akan diolah menggunakan algoritma machine learning Naïve Bayes. Setelah diproses klasifikasi dengan Naïve Bayes, maka akan keluar hasil dari kinerja pada model ini. Gambar 2 merupakan kerangka penelitian yang digunakan.

2.7. Tahapan Penelitian

Tahapan penelitian yang dilakukan dalam penelitian ini dapat dilihat pada Gambar 3.



Gambar 3. Tahapan Penelitian

Pada Gambar 3, tahapan yang dilakukan dalam penelitian ini sebagai berikut:

1. Pengumpulan Data

Pengumpulan data dilakukan dengan mengunduh dataset diabetes di UCI Repository; dataset ini dapat diunduh di <https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset>.

2. Model yang diusulkan

Model yang diusulkan dalam penelitian ini adalah menggunakan algoritma klasifikasi Naïve Bayes dan seleksi fitur Greedy Forward Selection, seleksi fitur digunakan untuk menseleksi fitur yang relevan sehingga dapat meningkatkan kinerja dari model yang digunakan.

3. Eksperimen dan Pengujian Model

Eksperimen dan pengujian dilakukan dengan beberapa tahapan sebagai berikut:

- Menyiapkan dataset untuk eksperimen.
- Mendesain arsitektur Naïve Bayes.
- Melakukan training dan testing terhadap model Naïve Bayes dan mencatat hasil kinerja dari model.
- Mendesain arsitektur Naïve Bayes dan Greedy Forward Selection.
- Melakukan training dan testing terhadap model Naïve Bayes dan Greedy Forward Selection, kemudian mencatat hasil kinerja dari model.

Spesifikasi komputer yang digunakan dalam eksperimen ini dapat ditunjukkan pada Tabel 3.

Tabel 3. Spesifikasi Komputer

Processor	AMD A9-9420 RADEON R5
RAM	4.00GB
Sistem Operasi	Windows 10
Aplikasi	Rapidminer

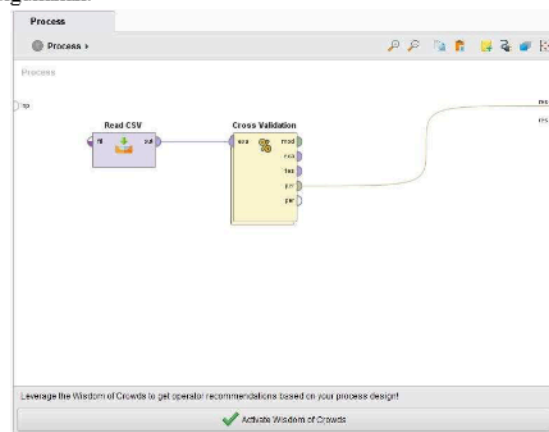
4. Evaluasi dan Validasi Hasil

Pada saat eksperimen dan pengujian selesai akan keluar hasil dari kinerja model yang diusulkan yaitu nilai akurasi sebagai ukuran seberapa bagus model yang digunakan.

3. HASIL

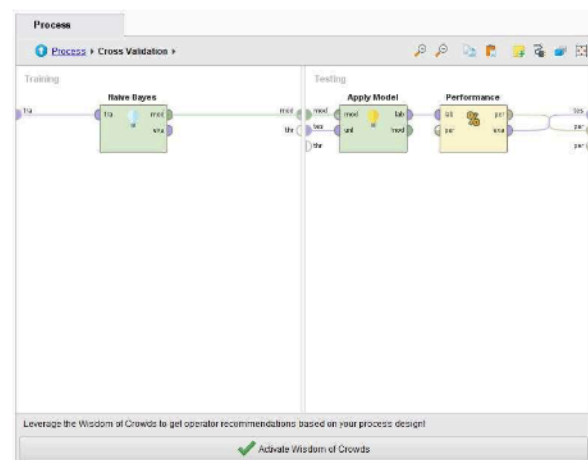
3.1. Eksperimen Naïve Bayes

Pada Gambar 4, dapat dilihat tampilan eksperimen di Rapidminer, dimana operator Read CSV digunakan untuk memanggil dataset diabetes, sedangkan Cross Validation digunakan untuk membagi dataset menjadi 10 bagian. Dataset yang dibagi menjadi 10 bagian, terdiri dari 2 jenis yaitu data training dan data testing. Data training fungsinya untuk melatih model sedangkan data testing untuk menguji model yang digunakan.



Gambar 4. Tampilan Rapidminer Memanggil Dataset

Gambar 5 menunjukkan eksperimen di Rapidminer dimana operator Naïve Bayes merupakan algoritma Naïve Bayes, sedangkan operator Apply Model digunakan untuk mengeluarkan rule dari Naïve Bayes dan operator Performance digunakan untuk mengeluarkan hasil kinerja atau akurasi dari model yang digunakan.



Gambar 5. Tampilan Rapidminer Mengolah Dataset

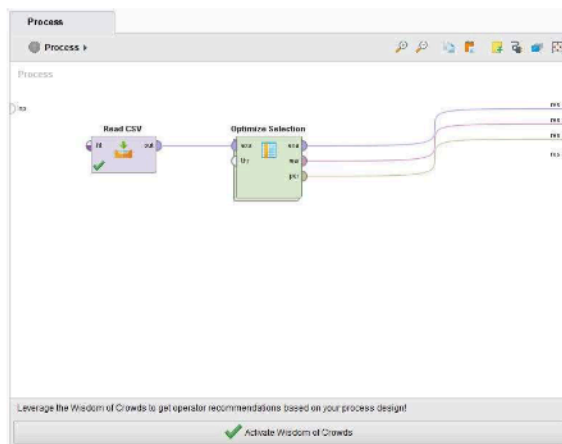
	True Positive	True Negative	Class Precision
pred. Positive	276	25	92.34%
pred. Negative	48	133	92.34%
class recall	86.25%	91.96%	

Gambar 6. Hasil Kinerja Model Naïve Bayes

Hasil kinerja atau hasil akurasi dapat dilihat pada Gambar 6, dimana hasil akurasi dari eksperimen di Rapidminer menggunakan algoritma Naïve Bayes adalah 87.69%.

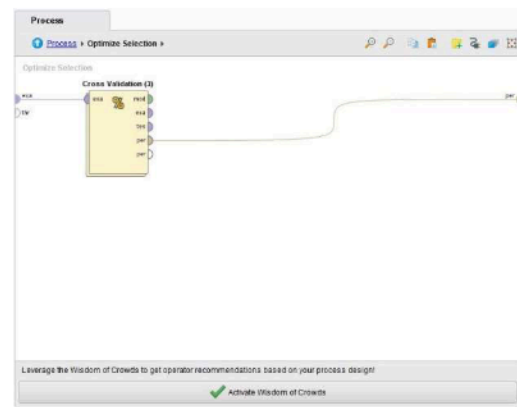
3.2. Eksperimen Naïve Bayes dan Greedy Forward Selection

Pada Gambar 7 dapat dilihat tampilan eksperimen di Rapidminer pada saat memanggil dataset menggunakan operator Read CSV. operator Optimize Selection merupakan operator untuk seleksi fitur menggunakan Greedy Forward Selection. Seleksi fitur digunakan untuk memilih fitur yang paling relevan.

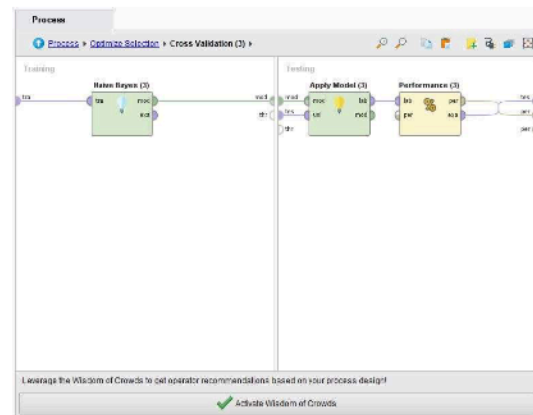


Gambar 7. Tampilan Rapidminer Dataset dan Seleksi Fitur

Gambar 8 menunjukkan operator Cross Validation, operator ini digunakan untuk membagi dataset menjadi 10 bagian secara otomatis, dimana bagian tersebut terdiri dari data training dan data testing.



Gambar 8. Tampilan Rapidminer Cross Validation



Gambar 9. Tampilan Rapiminer Mengolah Dataset

	True Positive	True Negative	Class Precision
pred. Positive	306	25	91.74%
pred. Negative	16	171	92.42%
class recall	95.62%	91.60%	

Gambar 10. Hasil Kinerja dari Model Naïve Bayes dan Greedy Forward Selection

Tampilan implementasi algoritma Naïve Bayes dapat dilihat pada Gambar 9, dimana operator Naïve Bayes digunakan. Opearator Apply model digunakan untuk menampilkan rule Naïve Bayes, sedangkan operator Performance digunakan untuk mengeluarkan hasil kinerja atau akurasi dari model yang digunakan. Pada Gambar 10 menunjukkan hasil kinerja atau akurasi dari model yang digunakan, dimana hasil akurasi model Naïve Bayes dan Greedy Forward Selection adalah 91.75%.

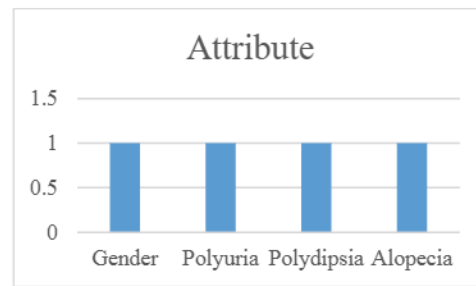
4. PEMBAHASAN

4.1. Hasil Seleksi Fitur Greedy Forward Selection

Seleksi fitur yang digunakan dalam penelitian ini adalah Greedy Forward Selection, diketahui bahwa seleksi fitur untuk memilih fitur-fitur yang paling relevan sehingga dapat meningkatkan kinerja dari model yang digunakan. Atribut yang berhasil diseleksi fitur menggunakan Greedy Forward Selection pada dataset diabetes adalah sebagai berikut:

1. Gender
2. Polyuria
3. Polydipsia
4. Alopecia

Gambar 11 dapat dilihat grafik atribut atau fitur yang dipilih menggunakan metode Greedy Forward Selection.



Gambar 11. Atribut Greedy Forward Selection

4.2. Hasil Naïve Bayes

Pada Tabel 4 dapat dilihat hasil dari confusion matrix algoritma Naïve Bayes tanpa menggunakan seleksi fitur Greedy Forward Selection, hasil ini berdasarkan eksperimen dan pengujian menggunakan aplikasi Rapidminer.

Tabel 4. Confusion Matrix Naïve Bayes

	True Positive	True Negative	Class Precision
Pred. Positive	276	20	93.24%
Pred. Negative	44	180	80.36%
Class Recall	86.25%	90.00%	

Keterangan:

TP (True Positive) = 276

TN (True Negative) = 180

FP (False Positive) = 20

FN (False Negative) = 44

$$Akurasi = \frac{276 + 180}{276 + 180 + 20 + 44} = 0.8769 = 87.69\%$$

$$Recall = TP_{rate} = \frac{276}{276 + 44} = 0.8625 = 86.25\%$$

$$Specificity = TN_{rate} = \frac{180}{180 + 20} = 0.9 = 90\%$$

$$Precision = PPV = \frac{276}{276 + 20} = 0.9324 = 93.24\%$$

$$NPV = \frac{180}{180 + 44} = 0.8035 = 80.35\%$$

4.3. Hasil Naïve Bayes dan Greedy Forward Selection

Pada Tabel 5 dapat dilihat hasil dari confusion matrix algoritma Naïve Bayes (NB) dan Greedy Forward Selection (GFS), hasil ini didapat dari eksperimen menggunakan aplikasi Rapidminer.

Tabel 5. Confusion Matrix NB+GFS

	True Positive	True Negative	Class Precision
Pred. Positive	306	29	91.34%
Pred. Negative	14	171	92.43%
Class Recall	95.62%	85.50%	

Keterangan:

TP (True Positive) = 306

TN (True Negative) = 171

FP (False Positive) = 29

FN (False Negative) = 14

$$Akurasi = \frac{306 + 171}{306 + 171 + 29 + 14} = 0.9173 = 91.73\%$$

$$Recall = TP_{rate} = \frac{306}{306 + 14} = 0.9562 = 95.62\%$$

$$Specificity = TN_{rate} = \frac{171}{171 + 29} = 0.855 = 85.5\%$$

$$Precision = PPV = \frac{306}{306 + 29} = 0.9134 = 91.34\%$$

$$NPV = \frac{171}{171 + 14} = 0.9243 = 92.43\%$$

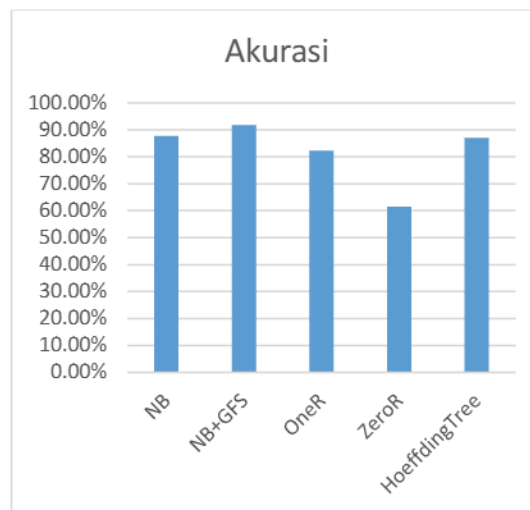
Berdasarkan hasil eksperimen dapat diketahui bahwa nilai kinerja terbaik itu dapat dilihat dari nilai akurasi. Nilai akurasi tertinggi ada pada algoritma Naïve Bayes dan Greedy Forward Selection (NB+GFS) sebesar 91.73%, sedangkan nilai akurasi untuk algoritma Naïve Bayes (NB) tanpa seleksi fitur Greedy Forward Selection adalah 87.69%. Dapat dilihat perbandingan nilai akurasi pada Tabel 6.

Tabel 6. Perbandingan Nilai Akurasi

Model	Akurasi
NB	87.69%
NB+GFS	91.73%

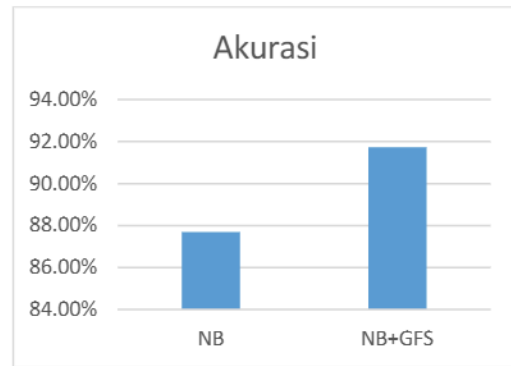
Tabel 7. Perbandingan Nilai Akurasi dengan Model Lain

Model	Akurasi
NB	87.69%
NB+GFS	91.73%
OneR	82.30%
ZeroR	61.53%
HoeffdingTree	87.11%



Gambar 13. Grafik Perbandingan Akurasi dengan Model Lain

Gambar 12 dapat dilihat perbandingan nilai akurasi menggunakan grafik sehingga sangat jelas terlihat perbedaanya.



Gambar 12. Grafik Perbandingan Nilai Akurasi

4.4. Perbandingan Hasil Penelitian

Perbandingan hasil penelitian algoritma Naïve Bayes (NB), Naïve Bayes dengan Greedy Forward Selection (NB+GFS), OneR, ZeroR dan HoeffdingTree dapat dilihat pada Tabel 7. Gambar 13 merupakan grafik perbandingan model NB, NB+GFS, OneR, ZeroR dan HoeffdingTree.

Tabel 8 merupakan perbandingan dengan penelitian yang lain diantaranya penelitian [11], hasil terbaik adalah algoritma K-Nearest Neighbor (KNN). Penelitian [12], hasil terbaik pada penelitian ini adalah Logistic Model Tree (LMT). Hasil terbaik algoritma Decision Tree dan Greedy Forward Selection

(DT+GFS) ada pada penelitian [10]. Support Vector Machine (SVM) merupakan hasil terbaik yang ada dalam penelitian [13]. Penelitian yang dibandingkan selanjutnya adalah penelitian [7], hasil terbaik ada pada algoritma Artificial Neural Network (ANN).

Tabel 8. Perbandingan Nilai Akurasi dengan Penelitian Lain

Model	Akurasi
NB	87.69%
NB+GFS	91.73%
KNN	100%
Logistic Model Tree	79.31%
DT+GFS	92.22%
SVM	82%
ANN	86.95%

5. KESIMPULAN

Diabetes merupakan penyakit kronis yang mematikan dan jika tidak diobati akan adanya komplikasi lain seperti penyakit stroke dan jantung. Dataset diabetes memiliki fitur-fitur yang tidak relevan sehingga dapat menurunkan kinerja dari algoritma yang digunakan, untuk menangani fitur-fitur yang tidak relevan diperlukan metode untuk pemilihan fitur yang relevan yaitu seleksi fitur atau atribut. Seleksi fitur ini dapat meningkatkan kinerja dari model yang digunakan, salah satunya adalah Greedy Forward Selection. Algoritma machine learning yang digunakan dalam penelitian ini adalah Naïve Bayes karena Naïve Bayes merupakan algoritma sederhana dan mudah. Hasil dari eksperimen yang dilakukan adalah nilai akurasi terbaik pada model Naïve Bayes dan Greedy Forward Selection (NB+GFS) sebesar 91.73%, sedangkan model Naïve Bayes mendapatkan nilai akurasi sebesar 87.69%. Dari eksperimen yang dilakukan dapat disimpulkan bahwa algoritma Naïve Bayes dan Greedy Forward Selection dapat memprediksi diabetes dengan sangat baik.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terimakasih M M Faniqul Islam, Rahatara Ferdousi, Sadikur Rahman, Humayra dan Yasmin Bushra atas dataset diabetes di UCI Repository.

DAFTAR PUSTAKA

[1] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, *Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques*, vol. 992. 2020.

[2] Noviandi, "Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes," *Inohim*, vol. 6, no. 1, pp. 1–5, 2018.

[3] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Icids, pp. 1578–1585, 2018.

[4] R. A. Siallagan and Fitriyani, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5," vol. 3, no. 1, pp. 45–46, 2021.

[5] R. S. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," *J. Softw. Eng.*, vol. 1, no. 1, 2015.

[6] C. Catal and B. Diri, "A systematic review of software fault prediction studies," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7346–7354, 2009.

[7] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, pp. 88–91, 2019.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[9] F. Fitriyani, "Metode Bagging Untuk Imbalance Class Pada Bedah Toraks Menggunakan Naive Bayes," *J. Kaji. Ilm.*, vol. 18, no. 3, p. 278, 2018.

[10] F. Fitriyani and T. Arifin, "Implementasi Greedy Forward Selection untuk Prediksi Metode Penyakit Kutil Menggunakan Decision Tree," *JST (Jurnal Sains dan Teknol.)*, vol. 9, no. 1, pp. 76–85, 2020.

- [11] S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *Int. J. Sci. Res.*, vol. 5, no. 1, pp. 1842–1845, 2016.
- [12] D. Vigneswari, N. K. Kumar, V. Ganesh Raj, A. Gagan, and S. R. Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus," *2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, pp. 84–87, 2019.
- [13] P. Sonar and K. Jaya Malini, "Diabetes prediction using different machine learning approaches," *Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019*, no. Iccmc, pp. 367–371, 2019.
- [14] V. Sindhu, S. A. S. Prabha, S. Veni, and M. Hemalatha, "Thoracic surgery analysis using data mining techniques," vol. 5, no. April, pp. 578–586, 2014.
- [15] R. Sanjaya and F. Fitriyani, "Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 316, 2019.
- [16] F. Fitriyani and R. S. Wahono, "Integrasi Bagging dan Greedy Forward Selection pada Prediksi Cacat Software dengan Menggunakan Naïve Bayes," *J. Softw. Eng.*, vol. 1, no. 2, pp. 101–108, 2015.
- [17] Alpaydm Ethem, *Introduction to Machine Learning Second Edition*, 2nd ed. London: MIT, 2010.
- [18] R. S. Wahono, N. S. Herman, and S. Ahmad, "Neural Network Parameter Optimization Based on Genetic Algorithm for Software Defect Prediction," vol. 20, no. 10, pp. 1951–1955, 2014.

BIODATA PENULIS



Fitriyani Memperoleh gelar S.T pada bidang Sistem Informasi di Universitas BSI Bandung dan gelar M.Kom pada bidang Ilmu Komputer di STMIK Nusa Mandiri. Minat penelitiannya adalah Data Mining, Machine Learning dan Software Engineering. Saat ini Dosen aktif di Universitas ARS.