Positif : Jurnal Sistem dan Teknologi Informasi
Volume 10, No 1, 2024, Page : 65 - 74

E-ISSN 2460-9552
P-ISSN 2620-3227

Article history
Received Sept 09, 2023
Accepted July 31, 2023
Published Nov 26, 2024

PERBANDINGAN ALGORITMA *SUPPORT VECTOR MACHINE* (SVM) DAN *NAÏVE BAYES* DALAM KLASIFIKASI PENYAKIT DIABETES

Anita Desiani[1*], Novi Rustiana Dewi[2], Muhammad Arhami[3], Dina Suzzete Sitorus[4], Suristhia Rahmadita[5]
Matematika dan Ilmu Pengetahuan Alam/Matematika, Universitas Sriwijaya[1*,2,4,5]
Teknologi informasi dan komputer, Poltek Lhokseumawe[3]
email: anita_desiani@unsri.ac.id[1*], novirustiana@unsri.ac.id[2],
muhammad.arhami@pnl.ac.id[3],suzzeteceliesitorus22@gmail.com[4], suristhiarahmadita08@gmail.com[5]

Abstract

High levels of sugar in the blood can cause diabetes. The longer people are unable to control glucose in their blood, the more complications it can cause, other diseases and even death. Early detection of diabetes is needed, one way is by carrying out data mining classification. Data mining classification in this research uses two algorithms, namely SVM (Support Vector Machine) and Naïve Bayes. This research compares the two algorithms using two methods, namely training split and k-fold cross validation which aims to get the best classification results in detecting diabetes. The best classification results are determined by calculating the average value of precision, recall and accuracy. Based on this research, the SVM algorithm with split percentage training produces average values for precision, recall and accuracy, namely 77%, 71.5%, 77.27%, while the SVM algorithm with k-fold cross validation produces average values for precision, recall , and accuracy is 77%, 72.5%, 71%. The Naïve Bayes algorithm with the split percentage training method produces average values for precision, recall and accuracy, namely 75.5%, 74.5%, 79%, while the Naïve Bayes algorithm with k-fold cross validation produces average values for precision, recall, and accuracy of 75.5%, 74.5%, 75%. The best classification result in detecting diabetes is the Naïve Bayes algorithm, the split percentage method, which provides the best accuracy, precision and recall values above 74%.

Keywords: *Classification, Diabetes, Comparison, Naïve Bayes, Support Vector Machine.*

Abstrak

Tingginya kadar gula dalam darah dapat mengakibatkan penyakit Diabetes. Semakin lama orang tidak dapat mengontrol glukosa dalam darah, maka dapat mengakibatkan komplikasi penyakit lain bahkan kematian. Diperlukannya deteksi dini terhadap penyakit diabetes, salah satu caranya yaitu dengan melakukan klasifikasi data mining. Klasifikasi data mining dalam penelitian ini menggunakan dua algoritma yaitu SVM (*Support Vector Machine*) dan Naïve Bayes. Penelitian ini membandingkan kedua algoritma dengan menggunakan dua metode yakni training split dan *k-fold cross validation* yang bertujuan untuk mendapatkan hasil klasifikasi terbaik dalam mendeteksi penyakit diabetes. Hasil klasifikiasi terbaik ditentukan dengan menghitung nilai dari rata-rata presisi, *recall*, dan akurasi. Berdasarkan penelitian ini, algoritma SVM dengan training *persentase split* menghasilkan nilai rata-rata untuk presisi, *recall*, dan akurasi yaitu 77%, 71.5%, 77.27%, sedangkan algoritma SVM dengan *k-fold cross validation* menghasilkan nilai rata-rata untuk presisi, *recall*, dan akurasi yaitu 77%, 72.5%, 71%. Algoritma *Naïve Bayes* dengan metode training persentase split menghasilkan nilai rata-rata untuk presisi, *recall*, dan akurasi yaitu 75.5%, 74.5%, 79%, sedangkan algoritma Naïve Bayes dengan *k-fold cross validation* menghasilkan nilai rata-rata untuk presisi, *recall*, dan akurasi sebesar 75.5%, 74.5%, 75%. Hasil klasifikasi terbaik dalam mendeteksi penyakit diabetes adalah algoritma Naïve Bayes metode persentase split memberikan nilai akurasi, presisi, *recall* terbaik diatas 74%.

Kata Kunci: Klasifikasi, Diabetes, Perbandingan, *Naïve Bayes, Support Vector Machine*

1. PENDAHULUAN

Hormon insulin merupakan rangkaian dari asam amino yang dihasilkan oleh kelenjar pankreas. Hormon insulin dibutuhkan oleh tubuh untuk mengatur keseimbangan kadar gula (glukosa) dalam darah [1]. Jika kekurangan hormon insulin, maka dapat menyebabkan kadar gula (glukosa) dalam darah tidak dapat dikontrol (tinggi) [2]. Tingginya kadar gula (glukosa) dalam darah mampu mengakibatkan penyakit *Diabets Millitus* atau sering dikenal dengan nama diabetes [3]. Penyakit diabetes yaitu salah satu jenis penyakit yang tidak menular namun tidak dapat disembuhkan. Semakin lama orang tidak dapat mengontrol glukosa (kadar gula) dalam darah dan menderita penyakit diabetes, maka dapat mengakibatkan penyakit komplikasi lain seperti kerusakan pada ginjal, saraf, kulit, permasalahan mata, dan bahkan dapat menyebabkan kematian [4].

Berdasarkan laporan Riskesdas pada tahun 2013 prevalensi diabetes mengalami peningkatan sebanyak 6,9% sedangkan pada tahun 2018 prevalensi diabetes meningkat sebanyak 8,5% dan dengan total kasus sebanyak 713.783 [5]. Pada tahun 2018 berdasarkan usia pravelensi diabetes yang paling sering terjadi pada rentang umur 15-24 tahun dengan jumlah 159.014 orang atau sebesar 22% [6]. Sedangkan pravelensi diabetes di Indonesia menduduki urutan ke tujuh di dunia pada tahun 2021 menurut IDF (*International Diabetes Federation)* banyak terjadi pada gologngan orang dewasa dengan rentang usia 20-79. Tingginya pravelensi diabetes di Indonesia terjadi karena terlalu banyak mengkonsumsi karbohidrat tinggi, jarang berolahraga, dan masih banyak lagi. Tingginya pravelensi diabetes di Indonesia, diperlukan deteksi dini terhadap penderita penyakit diabetes. Salah satu cara mendeteksi penyakit diabetes yaitu dengan memanfaatkan *data mining*, sehingga dapat membantu penderita penyakit diabetes untuk melakukan pengobatan yang relevan. *Data mining* adalah algoritma optimal data yang memiliki tujuan untuk mendapatkan informasi dari sekumpulan data jumlah tertentu [7]. Salah satu pengolahan *data mining* yaitu melakukan klasifikasi secara matematika. Klasifikasi merupakan sebuah proses dalam memasukan nilai sebuah objek data untuk masuk ke kelas tertentu sesuai dengan jumlah kelas yang ada [8]. Beberapa algoritma yang dapat digunakan *data mining* dalam klasifikasi adalah SVM (*Support Vector Machine*) dan *Naïve Bayes*. Pemanfaatan *data mining* dalam klasifikasi sudah banyak dilakukan dalam beberapa penelitian sebelumnya diantaranya; Damuri *et all* [9] melakukan klasifikasi pada dataset penerima bantuan sembako dengan menggunakan *Naïve Bayes* dengan akurasi 86%, Darmawan *et all* [10] melakukan klasifikasi pada dataset kepuasan pengunjung taman tabebuya dengan menggunakan algoritma SVM dengan akurasi 86% dan Arifin dan Sasangko [11] melakukan klasifikasi pada dataset jalur minat anak SMA dengan menggunakan SVM dan *Naïve Bayes* menghasilkan akurasi 97% dan 92%.

Algoritma SVM (*Support Vector Machine*) merupakan algoritma yang mampu bekerja dalam menentukan titik maksimal dan garis pemisah yang terbaik untuk memisah dua buah kelas [12]. Algoritma SVM memiliki kelebihan yaitu dapat menemukan *hyperplane* yang optimal berfungsi sebagai pemisah yang nyata bagi titik-titik masukan, mampu bekerja dan menghasilkan nilai terbaik pada *dataset* yang kecil, dan sangat cocok untuk data berdimensi tinggi [13], [14]. Penelitian yang menggunakan algoritma SVM diantaranya; Hermanto *et all* [15] menerapkan algoritma SVM untuk klasifikasi layanan komplain mahasiswa dengan akurasi 92,2% dan Dharmapatni dan Merawati [16] menerapkan algoritma SVM untuk klasifikasi kenaikan tarif BPJS kesehatan dengan akurasi 92%. Algoritma SVM juga memiliki kelemahan yaitu pemilihan kernel yang dapat mempengaruhi kinerja SVM, tidak cocok untuk *dataset* yang besar, tidak cocok untuk kumpulan data dengan nilai yang hilang. Berbeda dengan algoritma *Naïve bayes* yang mampu menangani data yang hilang dan cocok untuk *dataset* yang besar dengan dimensi tinggi karena pada *Naïve Bayes* perhitungannya menggunakan probabilitas untuk data yang hilang akan ditambahkan angka 1 pada setiap perhitungannya. Selain itu kelebihan *Naïve Bayes* yaitu performa klasifikasinya tinggi dan efisien karena tidak menggunakan kernel, regulasi, dan lain-lainnya [17]. *Naïve Bayes* merupakan algoritma pengklasifikasian statistik yang menggunakan probabilistik sederhana dan menerapkan teorema bayes (aturan bayes) dengan dugaan kuat yang tidak terikat [18]. Penelitian sebelumnya yang menggunakan *Naïve Bayes* diantaranya; Buani [19] yang memakai algoritma *Naïve Bayes* untuk klasifikasi pendeteksi gagal jantung dengan akurasi 70% dan Damuri *et all* [9]menggunakan algoritma *Naïve Bayes* untuk klasifikasi kelayakan penerima bantuan sembako dengan akurasi 86%, serta Desiani [20] yang

E-ISSN 2460-9552
P-ISSN 2620-3227

menggunakan algoritma *Naïve Bayes* untuk klasifikasi penyakit hati dengan akurasi 85%. Kelemahan algoritma *Naïve Bayes* yaitu probabilitas kurang berjalan secara optimal dan tidak cocok untuk tipe data numerik [19].

Penelitian ini, akan membandingkan dua algoritma yaitu SVM (*Support Vector Machine*) dan *Naïve Bayes* yang bertujuan mendapatkan hasil klasifikasi yang terbaik dalam mendeteksi penyakit diabetes. Dataset pada penelitian ini menghasilkan dua kelas yaitu diabetes dan tidak diabetes. Dalam studi saat ini, ukuran akurasi, recall, dan presisi akan diambil. Untuk pengujian kedua algoritma, desain pengujian pemisahan persentase dan desain validasi silang k-lipat digunakan. Untuk metode pemisahan persentase, ukuran pemisahan ditetapkan pada 80% untuk data uji dan 20% untuk data pelatihan. Untuk validasi silang k-lipat, nilai k diambil sebagai 10, yang berarti bahwa data dibagi menjadi sepuluh kelompok yang bergiliran sebagai data pelatihan dan pengujian selama sepuluh kali total. Hasilnya akan dibandingkan satu sama lain agar mendapatkan algoritma dan metode uji yang terbaik yang dapat digunakan dalam mendeteksi penyakit diabetes.

2. METODE PENELITIAN

Deskripsi Data

Data set yang digunakan dalam penelitian ini adalah data set yang berasal dari situs Kaggle (www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset) dengan judul Diabetes Dataset dan berformat csv. Data set tersebut diperoleh dari National Institute of Diabetes and Digestive and Kidney Diseases. Secara total, jumlah data berjumlah 768 perempuan Pima Indian berusia di atas 21 tahun. Pada dataset tersebut terdapat 9 variabel yang terdiri dari 8 variabel independent dan 1 variabel dependent. Variabel yang bersifat independent adalah pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, dan age. Variabel dependent dalam penelitian ini adalah label. Terdapat 2 kelas label pada dataset ini, kelas 0 tidak diabetes berjumlah 500, kelas 1 diabetes berjumlah 258. Selamat, hanya label pada 2 kelas. Klasifikasi label dibuat pada faktor transformasi dependent dan konfigurasinya memberikan hasil yang ditunjukkan pada tabel berikut. Mari melihat tabel 1 di atas yang memuat catatan dari variabel-variabel yang digunakannya pada dataset diabetes.

Tabel 1. Variabel Data

Atribut	Keterangan	Deskripsi
Pregnancies	Kehamilan	Jumlah berapa kali wanita hamil selama hidupnya
Glucose	Kadar Gula	Konsentrasi glukosa pada 2 jam dalam tes toleransi glukosa
Blood Pressure	Tekanan Darah	Tekanan darah diastolik dalam (mm/Hg) ketika jantung rileks setelah kontraksi
Skin Thickness	Penebalan Kulit	Memperkirakan lemak tubuh (mm) yang diukur pada lengan kanan.
Insulin	Insulin	Tingkat insulin 2 jam insulin serum dalam satuan mu U/ml
BMI	Indeks Massa Tubuh	Berat dalam kg / (tinggi dalam meter kuadrat), dan merupakan indikator kesehatan seseorang
Diabetes Pedigree Function	Fungsi Silsilah Diabetes	Indikator riwayat diabetes dalam keluarga
Age	Umur	Umur wanita suku Indian Pima
Outcome	Label	0 = Tidak diabetes, 1 = Diabetes

Preprocessing Data

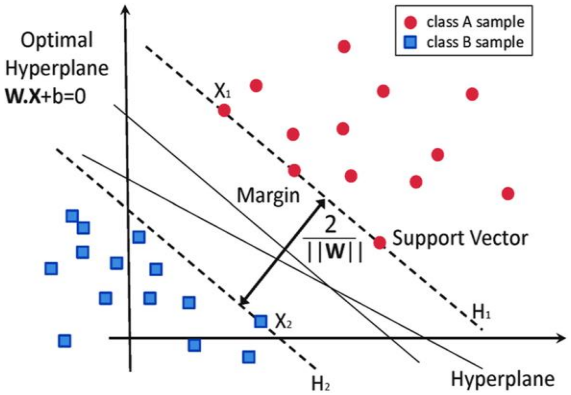
Preprocessing data adalah teknik awal dari *data mining* untuk memperbaiki data mentah menjadi informasi yang mudah dipahami. *Preprocessing* data atau praproses data dilakukan untuk mempermudah proses analisis data, mengurangi durasi *data mining*, dan mendapatkan hasil yang lebih tepat [21]. Praproses data dapat berupa transformasi data. Praproses data pada dataset diabetes akan ditransformasikan karena pada atribut label akan diubah tipe data *numeric* menjadi kategorik khusus untuk algoritma *Naïve Bayes*. Dataset diabetes akan ditransformasikan karena ada rentang nilai yang memiliki perbedaan di beberapa atribut. Untuk mengatasinya maka dilakukan sebuah transformasi data menggunakan normalisasi dengan tujuan agar rentang nilai setiap atribut sama, sehingga dapat memperoleh klasifikasi yang lebih baik. Dalam penelitian ini, terdapat 8 atribut yang ditransformasi dengan normalisasi. Pada atribut Kehamilan, Kadar Gula, Tekanan Darah, Penebalan Kulit, *Insulin,* Indeks Massa Tubuh, Fungsi Silsilah Diabetes dan *Umur* karena di dalamnya terdapat range data yang terlalu jauh. Normalisasi data yang dipakai dalam penelitian ini yaitu dengan melakukan teknik *min-max normalization* dengan persamaan (1) berikut [21].

$$normalized\ (x) = \frac{Range\ x + (x\ minValue)(maxRange - minRange)}{maxValue - minValue} \quad (1)$$

Untuk memisahkan klasifikasi ini, digunakan 2 metode pengujian, yaitu persentase bersplit dan k-fold cross validation. Pada persen split ini memiliki perbandingan 8:2 yang artinya, data tersebut akan terpecah menjadi 2 bagian yakni data training dan data testing. Jika data diletakkan pada dryink pens atau dryink.nama, data ini bersifat tetap yang hanya akan diverifikasi dalam pengujian. 80 persen dari data akan digunakan sebagai data pelatihan, dan 20 persen akan digunakan sebagai data evaluasi. Menggunakan k=10 untuk k-fold cross validation.

Algoritma Support Vector Machine (SVM)

SVM digunakan untuk menemukan *hyperplane* optimal sehingga dapat memisahkan ke dalam dua kelas yang berbeda serta memaksimalkan margin antara dua kelas tersebut [12]. Tujuan dari algoritma SVM dapat dilihat pada Gambar 1.



Gambar 1. *Hyperplane* Terbaik Untuk Memisahkan Kedua Kelas

Gambar 1. merupakan garis *hyperplane* terbaik yang dapat memisahkan dua kelas yang disimbolkan dengan kotak biru dan lingkaran merah. Garis-garis terputus merupakan simbol dari bidang pembatas yang memisahkan dua kelas tersebut secara sejajar, sehingga didapatkan pertidaksamaan (2) dan pertidaksamaan (3) , yaitu [22], [23].

$$\vec{w}.\vec{x}_i + b \geq 1, untuk\ kelas\ 1 \quad (2)$$
$$\vec{w}.\vec{x}_i + b \geq 0, untuk\ kelas\ 0 \quad (3)$$

$\vec{w}$ merupakan normal bidang atau nilai bobot, b adalah nilai bias atau posisi bidang *alternative* terhadap pusat koordinat, dan $\vec{x}_i$ merupakan nilai input ke-i dimana $\vec{x}_i \in \overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_n}$. *Hyperplane* optimal memiliki marginal, seperti pada Gambar 1. Marginal didapatkan dengan memaksimalkan jarak antara *hyperplane* dan titik kedua kelas terdekat. Rumus mencari marginal adalah $\frac{1-b(-1-b)}{\|\vec{w}\|} = \frac{2}{\|\vec{w}\|}$. Memaksimalkan marginal dengan tetap memenuhi pertidaksamaan (2) dan (3). Apabila kedua batasan bidang pada ketidaksamaan (2) dan (3) akan direpresentasikan dalam ketidaksamaan (4), yaitu

$$\min \frac{1}{2}\|\vec{w}\|^2 \text{ dengan } y_i(\vec{w}.\vec{x_i} + b) - 1 \geq 0 \quad (4)$$

y_i merupakan kelas label ke-i sampai N. Untuk memperoleh pengklasifikasian seperti pada Gambar 1, maka pengklasifikasian data tidak dapat dipisahkan secara linear. Formula SVM harus dimodifikasi dengan menggunakan pertidaksamaan (4) dan penambahan variabel ξ_i

E-ISSN 2460-9552
P-ISSN 2620-3227

dimana $\xi_i \geq 1, \forall_i : \xi_i$ didapatkan pertidaksamaan (5)

$$\min \frac{1}{2}\|\vec{w}\|^2 + C\left(\sum_{i=1}^{N} \xi_i\right) \text{ dengan } y_i(\vec{w}.\vec{x_l} + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \qquad (5)$$

C adalah parameter yang digunakan untuk mengurangi kesalahan data pelatihan sambil menyederhanakan model, C juga dikenal sebagai parameter regularisasi. Untuk penelitian ini, fungsi kernel yang digunakan adalah kernel linier,, yaitu $K(\vec{x}.\vec{y}) = \vec{x}.\vec{y}$. $K(\vec{x}.\vec{y})$ adalah fungsi kernel linear, $\vec{x}$ adalah data latih (training) dan $\vec{y}$ adalah data uji.

Algoritma Naïve Bayes

Naïve Bayes adalah metode pengklasifikasian statistika dengan menggunakan probabilistik yang sederhana dan menerapkan teorema bayes (aturan bayes) dengan dugaan kuat yang tidak terikat [18]. Segala atribut diperlakukan secara bebas dan sama antara satu atribut dengan atribut yang lainnya. Metode ini memakai *Naïve Bayes Classifier* dalam perhitungan nilai bobot peluang pada setiap atribut. Langkah-langkah yang digunakan dalam klasifikasi dengan metode *Naïve Bayes* sebagai berikut [24]:

1. Dihitung jumlah kelas atau label;
2. Setiap kategori akan dihitung peluang;
3. Menentukan jumlah kemunculan atau frekuensi untuk setiap kategori;
4. Dengan nilai maksmimal akan ditentukan kategorinya.

Perhitungan algoritma *Naïve Bayes* dilakukan dengan menggunakan persamaan (13) berikut:

$$P(M) = \frac{P(M|N)\,P(N)}{P(N)} \qquad (13)$$

Untuk M adalah data yang kelasnya belum diketahui. N adalah hipotesis data M merupakan kelas spesifik. P merupakan simbol peluang, maka $P(N)$ itu peluang hipotesis $N, P(M)$ itu peluang hipotesis M. Dan untuk simbol | merupakan peluang bersyarat, maka $P(M|N)$ merupakan peluang hipotesis N berdasarkan kondisi M sedangkan $P(N|M)$ merupakan peluang hipotesis M berdasarkan kondisi N.

Analisis Hasil

Hasil disajikan menggunakan *confusion matrix. Confusion matrix* adalah tabel yang memberikan informasi tentang jumlah total titik data yang diuji di luar sampel yang telah diklasifikasikan dengan benar dan uji di luar sampel yang telah diklasifikasikan secara salah biasanya dengan angka dari total tes yang dilakukan [17]. Dari segi lain, definisi lain dari *confusion matrix* adalah matriks yang menggambarkan kemampuan dari klasifikasi algoritma dengan mesin berbentuk matriks yang mengklasifikasikan hasil prediksi menjadi empat kategori, yaitu: *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Ciri-ciri dan pembentukan *confusion matriks* untuk klasifikasi dua kelas dapat dilihat pada Tabel 2 [18].

Tabel 2. *Confusion Matrix*

Kelas		**Nilai Aktual**	
		Positif	Negatif
Nilai Prediksi	Positif	True Positive (TP)	False Positive (FP)
	Negatif	False Negative (FN)	True Negative (FN)

Keterangan [25]:

1. *True Positive (TP)* adalah total data positif yang diklasifikasikan sebagai positif.
2. *False Negative (FN)* adalah total data negatif yang diklasifikasikan sebagai positif.
3. *False Positive (FP)* adalah total data positif yang diklasifikasikan sebagai negatif.
4. *True Negative (TN)* adalah total data negatif yang diklasifikasikan sebagai negatif.

Dari *confusion matrix*, kita dapat menghitung akurasi, presisi, dan recall. Akurasi di sisi lain dapat ditentukan oleh seberapa akurat klasifikasi tersebut. Akurasi dari suatu klasifikasi dapat mencerminkan kinerja keseluruhan dari model klasifikasi. Dengan tingkat akurasi yang tinggi, dapat disimpulkan bahwa model klasifikasi tersebut akan memiliki kinerja yang baik, sedangkan tingkat akurasi yang rendah menyiratkan kinerja yang buruk dari model klasifikasi. Rumus yang dapat digunakan untuk menentukan akurasi ditunjukkan dalam persamaan (14) [26]:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \text{ x } 100\% \qquad (14)$$

Presisi ialah besaran nilai ketepatan antara informasi yang diinginkan oleh pengguna terhadatanggapan yang diberikan oleh sistem. Rumus dalam menghitung presisi dapat dilihat pada persamaan (15) [27]:

$$\text{Presisi} = \frac{TP}{TP + FP} \text{ x } 100\% \qquad (15)$$

Recall ialah besaran nilai ketepatan sistem dalam mendapatkan kembali sebuah penjelasan informasi. Rumus yang digunakan dalam menghitung *recall* dapat dilihat pada persamaan (16) [26]:

$$\textit{Recall} = \frac{TP}{TP + FN} \text{ x } 100\% \qquad (16)$$

3. HASIL DAN PEMBAHASAN

Hasil Algoritma SVM

Penerapan algoritma SVM pada penyakit diabetes. *Confusion matriks* dari algoritma SVM pada dataset penyakit diabetes dengan metode *training persentase split* dan *k-fold validation* terlihat pada Tabel 3.

Tabel 3. *Confusion Matrix SVM (Support Vector Machine)*

Confusion Matrix Persentase Split dan K-Fold Cross Validation			
Kelas		Nilai Aktual	
		Diabetes	Tidak Diabetes
Nilai Prediksi dengan Persentase Split	Diabetes	91	9
	Tidak Diabetes	26	28
Nilai Prediksi dengan Persentase K-Fold Cross Validation	Diabetes	415	85
	Tidak Diabetes	115	153

Berdasarkan Tabel 3 dapat dilihat bahwa *persentase split* memprediksi 91 pasien diabetes sebagai diabetes, 26 pasien diabetes sebagai tidak diabetes, 9 pasien tidak diabetes sebagai diabetes, dan 28 pasien diabetes sebagai diabetes. Algoritma SVM dengan metode *k-fold cross validation* memprediksi 119 secara tepat dan 35 data ditebak pada kelas yang salah. Selanjutnya dihitung nilai presisi, *recall*, dan akurasi terlihat pada Tabel 4.

Tabel 4. Perbandingan kedua teknik pengujian *algoritma SVM (Support Vector Machine)*

Metode *Training*	Akurasi	Presisi		Recall	
		Diabetes	Tidak Diabetes	Diabetes	Tidak Diabetes
Persentase Split	77%	76%	78%	52%	91%
K-Fold Cross Validation	71%	76%	78%	54%	91%

Berdasarkan tabel 4 terlihat bahwa nilai akurasi, presisi, dan *recall* yang diperoleh dari penerapan algoritma SVM menggunakan teknik pengujian persentase split lebih besar dibandingan dengan teknik pengujian *k-fold cross validation.* Nilai akurasi sebesar 77% dengan presisi diabetes 76% dan tidak diabetes 78%. Untuk nilai *recall* diabetes dan tidak diabetes sebesar 52% dan 91%.

Hasil Algoritma Naïve Bayes

Penerapan algoritma *Naïve Bayes* pada penyakit diabetes *confusion matriks* dari algoritma *Naïve Bayes* pada dataset penyakit diabetes dengan metode *training persentase split* dan *k-fold validation* terlihat di Tabel 5.

Tabel 5. *Confusion Matrix Naïve Bayes*

Confusion Matrix Persentase Split dan K-Fold Cross Validation			
Kelas		Nilai Aktual	
		Diabetes	Tidak Diabetes
Nilai Prediksi dengan Persentase Split	Diabetes	93	14
	Tidak Diabetes	18	29
Nilai Prediksi dengan Persentase *K-Fold Cross Validation*	Diabetes	422	78
	Tidak Diabetes	109	159

Dari Tabel 5 dapat dilihat bahwa *Persentase Split* memprediksi 93 pasien diabetes sebagai diabetes, 18 pasien diabetes sebagai tidak diabetes, 14 pasien tidak diabetes sebagai diabetes, dan 29 pasien diabetes sebagai diabetes. Kemudian untuk

E-ISSN 2460-9552
P-ISSN 2620-3227

k-fold validation memprediksi 93 pasien diabetes sebagai diabetes, 19 pasien diabetes sebagai tidak diabetes, 14 pasien tidak diabetes sebagai diabetes, dan 28 pasien diabetes sebagai diabetes. Metode *Naïve Bayes* memprediksi 121 secara tepat dan 33 data ditebak pada kelas salah. Selanjutnya dihitung nilai presisi, recall, dan akurasi terlihat pada Tabel 6.

Tabel 6. Perbandingan kedua teknik pengujian *algoritma Naïve Bayes*

Metode Training	Akurasi	Presisi		Recall	
		Diabetes	Tidak Diabetes	Diabetes	Tidak Diabetes
Persentase Split	79%	67%	84%	62%	87%
K-Fold Cross Validation	75%	71%	80%	64%	85%

Berdasarkan Tabel 6 terlihat bahwa nilai akurasi, presisi, dan *recall* yang diperoleh dari penerapan algoritma *Naïve Bayes* menggunakan teknik pengujian *Persentase Split* lebih besar dibandingan dengan teknik pengujian *k-fold cross validation.* Nilai akurasi sebesar 79% dengan presisi diabetes 67% dan tidak diabetes 84%. Untuk nilai *recall* diabetes dan tidak diabetes sebesar 62% dan 87%.

Hasil Algoritma C4.5

Penerapan algoritma C4.5 pada penyakit diabetes *confusion matriks* dari algoritma C4.5 pada dataset penyakit diabetes dengan metode *training persentase split* dan *k-fold validation* terlihat di Tabel 7.

Tabel 7. *Confusion Matrix* C4.5

Confusion Matrix Persentase Split dan K-Fold Cross Validation			
Kelas		Nilai Aktual	
		Diabetes	Tidak Diabetes
Nilai Prediksi dengan Persentase Split	Diabetes	77	22
	Tidak Diabetes	25	30
Nilai Prediksi dengan Persentase *K-Fold Cross Validation*	Diabetes	396	104
	Tidak Diabetes	112	156

Dari Tabel 7 dapat dilihat bahwa *Persentase Split* memprediksi 77 pasien diabetes sebagai diabetes, 25 pasien diabetes sebagai tidak diabetes, 22 pasien tidak diabetes sebagai diabetes, dan 30 pasien tidak diabetes sebagai tidak diabetes. Kemudian untuk *k-fold validation* memprediksi 396 pasien diabetes sebagai diabetes, 112 pasien diabetes sebagai tidak diabetes, 104 pasien tidak diabetes sebagai diabetes, dan 28 pasien tidak diabetes sebagai tidak diabetes. Metode C4.5 memprediksi 107 secara tepat dan 47 data ditebak pada kelas salah. Selanjutnya dihitung nilai presisi, recall, dan akurasi terlihat pada Tabel 8.

Tabel 8. Perbandingan kedua teknik pengujian *algoritma C4.5*

Metode Training	Akurasi	Presisi		Recall	
		Diabetes	Tidak Diabetes	Diabetes	Tidak Diabetes
Persentase Split	69%	58%	75%	55%	78%
K-Fold Cross Validation	71%	60%	78%	58%	79%

Berdasarkan Tabel 8 terlihat bahwa nilai akurasi, presisi, dan *recall* yang diperoleh dari penerapan algoritma C4.5 menggunakan teknik pengujian *k-fold cross validation* lebih besar dibandingan dengan teknik pengujian *Persentase Split.* Nilai akurasi sebesar 71% dengan presisi diabetes 58% dan tidak diabetes 75%. Untuk nilai *recall* diabetes dan tidak diabetes sebesar 55% dan 78%.

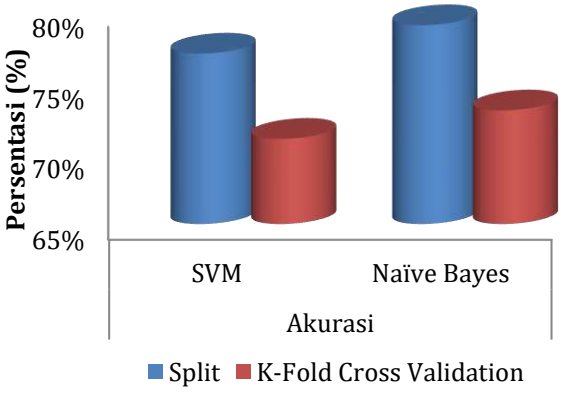
Hasil Perbandingan SVM dan Naïve Bayes

Hasil prediksi dari dua algoritma SVM dan *Naïve Bayes* terlihat bahwa untuk mengklasifikasikan penyakit diabetes metode SVM dan *Naïve Bayes* dapat bekerja dengan baik karena diatas 70%. Setelah dilakukan perhitungan nilai akurasi, presisi, dan *recall* dengan algoritma SVM dan *Naïve Bayes* memakai teknik *training presentase* split dan *k-fold cross validation* dapat dibandingkan hasil keduanya, dapat dilihat pada Tabel 9.

Tabel 9. Perbandingan *SVM dan Naïve Bayes* dengan *presentase split dan K-Fold Cross Validation*

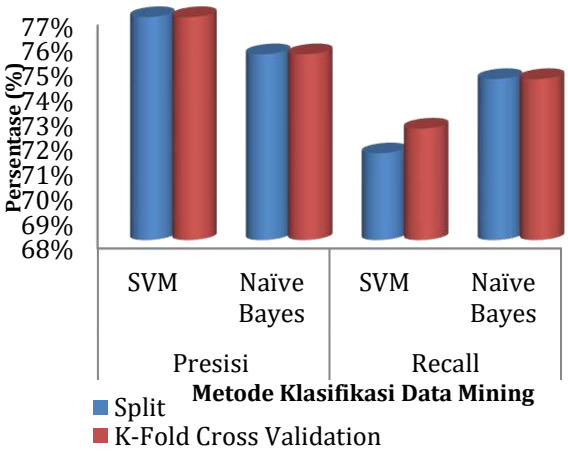
Metode Training	*Presentase Split*		*K-Fold Cross Validation*	
	SVM	*Naïve Bayes*	SVM	*Naïve Bayes*
Akurasi	77.27%	79%	71%	75%
Presisi	77%	75.5%	77%	75.5%
Recall	71.5%	74.5%	72.5%	74.5%

Dengan mengacu pada Tabel 9 data latihan pada rasio pembagian algoritma *Naïve Bayes* menunjukkan nilai presisi, *recall*, dan akurasi yang lebih baik dibandingkan algoritma SVM yang nilainya berturut-turut 75.5, 74.5 dan 79%. Sedangkan untuk algoritma SVM memperoleh nilai presisi, recall, dan akurasi masing-masing berturut 77%, 71.5%% dan 77.27%. Untuk teknik pengujian k-fold cross validation juga didapatkan hasil *Naïve Bayes* dengan nilai akurasi, precision dan recall lebih tinggi dari pada algoritma SVM dengan rasio berturut-turut 75% dan 75.5% dan 74.5%. Sementara itu algoritma SVM hanya mampu meraih nilai akurasi, presisi dan recall masing-masing berturut 71%, 77% dan 72,5%. Dengan ini disimpulkan bahwa berdasarkan hasil kedua teknik pengujian algoritma *Naïve Bayes* memiliki peringkat akurasi, presisi dan *recal*l yang lebih baik dibandingkan algoritma SVM. Presentasi penyebaran metode pengujian *split* lebih baik daripada metode *k-fold cross validation* pada semua data yang diujikan di kedua algoritma klasifikasi. Untuk mempermudah dalam membaca hasil perbandingan nilai akurasi dari kedua algoritma terlihat pada gambar2.



Gambar 2. Akurasi SVM dan *Naïve Bayes* Menggunakan Presentase Split dan *K-Fold Cross Validation*

Berdasarkan Gambar 2 dapat dilihat nilai akurasi dari algoritma SVM dan *Naïve Bayes*, besar nilai akurasi yang dihasilkan *persentase split* lebih besar dibandingkan *k-fold cross validation.* Pada algoritma *Naïve Bayes* nilai akurasi dengan *persentase* split lebih besar dibandingkan SVM yaitu 79% sedangkan SVM 77%. Kemudian rata-rata nilai presisi dan *recall* pada metode SVM dan *Naïve Bayes* terlihat pada Gambar 3.



Gambar 3. Nilai Rata-rata Presisi dan *Recall* Algoritma SVM dan *Naïve Bayes*

Berdasarkan Gambar 3 dapat diketahui bahwa SVM pada algoritma yang menggunakan metode persentase *split* dan *k-fold cross validation* memiliki tingkat presisi sebesar 77% yang lebih tinggi dibandingkan *Naïve Bayes* sebesar 75.5%. Sementara untuk *average recall* pada *Naïve Bayes* dengan persentase *split* dan *k-fold cross validation* adalah sebesar 74.5% yang lebih besar dibandingkan pada algoritma SVM. Untuk nilai rata-rata recall, nilai rata-rata *k-fold cross validation* sebesar 72.5% lebih besar dibandingkan dengan persentase *split* yang berformat 71.5%. Dan Metode *Naïve Bayes* dengan persentase *split* melebihi *Naïve Bayes* dengan *k-fold cross*. *Naïve Bayes* dengan persentase *spli*t, SVM dengan persentase *split*, dan SVM dengan persentase *k-fold* juga menggunakan metode *cross validation.*

4. PENUTUP

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan dapat disimpulkan bahwa klasifikasi penyakit diabetes dengan menggunakan metode *Support Vector Machine* (SVM) dan metode *Naïve Bayes* tergolong baik

E-ISSN 2460-9552
P-ISSN 2620-3227

karena diatas 70%. Hal ini dapat dilihat dari nilai akurasi, presisi, dan recall dari kedua metode. Metode SVM dengan metode uji yaitu *persentase split* menghasilkan nilai rata-rata untuk presisi, recall, dan akurasi yaitu 77%, 71.5%, 77.27%, sedangkan metode SVM dengan *k-fold cross validation* menghasilkan nilai rata-rata untuk presisi, *recall*, dan akurasi yaitu 77%, 72.5%, 71%. Metode *Naïve Bayes* dengan metode training persentase split menghasilkan nilai rata-rata untuk presisi, *recall*, dan akurasi yaitu 75.5%, 74.5%, 79%, sedangkan metode *Naïve Bayes* dengan *k-fold cross validation* menghasilkan nilai rata-rata untuk presisi, *recall*, dan akurasi sebesar 75.5%, 74.5%, 75%. Hasil klasifikasi terbaik dalam mendeteksi penyakit diabetes adalah metode *Naïve Bayes* dengan metode uji persentase split yang memberikan nilai akurasi, presisi, recall terbaik diatas 74%.

5. REFERENSI

[1] C. S. Sari, "Evaluasi Pemberian Informasi Obat Insulin Pada Pasien Rawat Jalan Rumah Sakit Pku Muhammadiyah Sekapuk," Universitas Muhammadiyah Gresik, 2020.

[2] U. Hasanah, "Insulin Sebagai Pengatur kadar Gula Darah," *J. Kel. Sehat Sejah.*, vol. 11, no. 22, pp. 42–49, 2013.

[3] D. N. Anisa and Jumanto, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Naive Bayes," *Din. Inform.*, vol. 14, no. 1, pp. 33–42, 2022.

[4] E. D. Nurcahya, "Klasifikasi Penyakit Ayam Menggunakan Metode Support Vector Machine," *VOLT J. Ilm. Pendidik. Tek. Elektro*, vol. 2, no. 1, p. 45, 2017.

[5] Kementrian Kesehatan Republik Indonesia, "Berat Badan Ideal Bantu Cegah Timbulnya Diabetes," *Biro Komunikasi dan Pelayanan Masyarakat*, 2021.

[6] Badan Penelitian dan Pengembangan Kesehatan, "Laporan Riskesdas 2018 Nasional," *Lembaga Penerbit Balitbangkes*. 2018.

[7] W. Hoaxiang and S. Smys, "Big Data Analysis and Perturbation Using Data Mining Algorithm," *J. Soft Comput. Paradig.*, vol. 3, no. 1, pp. 19–28, 2021.

[8] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020.

[9] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," *JURIKOM (Jurnal Ris. Komputer)*, vol. 8, no. 6, p. 219, 2021.

[10] A. Darmawan, N. Kustian, and W. Rahayu, "Implementasi Data Mining Menggunakan Model SVM Untuk Prediksi Kepuasan Pengunjung Taman Tabebuya," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 2, no. 3, pp. 299–307, 2018.

[11] O. Arifin and T. B. Sasongko, "Analisa Perbandingan Tingkat Performansi Metode Support Vector Machine dan Naive Bayes Classifier untuk Klasifikasi Jalur Minat SMA," *Semin. Nas. Teknol. Inf. dan Multimed. 2018*, pp. 67–72, 2018.

[12] A. Budianto, R. Ariyuana, and D. Maryono, "Perbandingan K-Nearest Neighbor (KNN) Dan Support Vector Machine (SVM) Dalam Pengenalan Karakter Plat Kendaraan Bermotor," *J. Ilm. Pendidik. Tek. dan Kejuru.*, vol. 11, no. 1, pp. 27–35, 2019.

[13] A. T. Novarina, E. Santoso, and Indriati, "Sistem Pakar Diagnosis Penyakit Hepatitis Menggunakan Metode Dempster Shafer," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 6, pp. 2252–2258, 2018.

[14] S. Sayed, M. Nassef, A. Badr, and I. Farag, "A Nested Genetic Algorithm For Feature Selection In High-Dimensional Cancer Microarray Datasets," *Expert Syst. Appl.*, vol. 121, pp. 233–243, 2019.

[15] H. Hermanto, A. Mustopa, A. Y. Kuntoro, and others, "Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa," *JITK (Jurnal Ilmu Pengetah. Dan Teknol. Komputer)*, vol. 5, no. 2, pp. 211–220, 2020.

[16] P. M. N. Dharmapatni and N. L. P. Merawati, "Penerapan Algoritma Support Vector Machine Dalam Sentimen Analisis

Terkait Kenaikan Tarif BPJS Kesehatan," *J. Bumigora Inf. Technol.*, vol. 2, no. 2, pp. 105–112, 2020.

[17] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique," *J. Med. Syst.*, vol. 43, no. 8, 2019.

[18] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019.

[19] D. Cahya Putri Buani, "Penerapan Algoritma Naïve Bayes dengan Seleksi Fitur Algoritma Genetika Untuk Prediksi Gagal Jantung," *EVOLUSI J. Sains dan Manaj.*, vol. 9, no. 2, pp. 43–48, 2021.

[20] A. Desiani, "Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati," *Simkom*, vol. 7, no. 2, pp. 104–110, 2022.

[21] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 4, no. 1, pp. 78–82, 2019.

[22] I. M. Parapat, M. T. Furqon, and Sutrisno, "Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10, pp. 3165–3166, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[23] D. A. Anggoro and N. D. Kurnia, "Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms In Predicting Heart Disease," *Int. J.*, vol. 8, no. 5, pp. 1689–1694, 2020.

[24] P. T. D. P. Putu, M. F. Zambak, Suwarno, and P. Harahap, "Analisa Radiasi Sinar Matahari Terhadap Panel Surya 50 WP," *RELE (Rekayasa Elektr. dan Energi) J. Tek. Elektro*, vol. 4, no. 1, pp. 48–54, 2021.

[25] A. Indriani, "Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier," *Semin. Nas. Apl. Teknol. Inf.*, pp. 1–10, 2014.

[26] B. P. Pratiwi, A. S. Handayani, and S. Sarjana, "Pengukuran Kinerja Sistem Kualitas Udara dengan Teknologi WSN Menggunakan Confusion Matrix," *J. Inform. Upgris*, vol. 6, no. 2, Jan. 2021.

[27] E. Prasetyo, *Data Mining: Konsep dan Aplikasi menggunakan MATLAB*. Yogyakarta: ANDI, 2012.