PERBANDINGAN KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE MACHINE LEARNING

Tasya Syamsudin [1] Teny Handhayani [2] Muhammad Isnaini Syaifudin [3]

[1] Teknik Informatika, FTI, Universitas Tarumanaraga
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
email : tasya.535200073@stu.untar.ac.id[1] tenyh@fti.untar.ac.id [2] muhammad.535200070@stu.untar.ac.id[3]

ABSTRAK

Diabetes adalah penyakit ketika tubuh manusia tidak dapat menggunakan insulin dengan baik. Apabila pada kasus tersebut berlangsung dalam waktu jangka panjang, maka kadar glukosa tersebut dapat merusak organ tubuh, bahkan kegagalan fungsi organ dan jaringan pada tubuh manusia yang dapat menyebabkan komplikasi bahkan kematian. Menurut International Diabetes Federation, pada tahun 2021, kematian yang disebabkan oleh diabetes sebanyak 236.711 ribu jiwa yang berusia sekitar 20-79 tahun. Perkembangan teknologi pada masa sekarang, dapat membantu manusia untuk mendapatkan informasi dan memprediksi penyakit tersebut serta dapat membantu dalam pengembangan pengobatan dan agar mencegah terjadinya penyakit diabetes tertentu lebih dalam menggunakan pendekatan machine learning dengan teknik klasifikasi. Algoritma klasifikasi yang akan digunakan penulis untuk memprediksi penyakit diabetes tersebut adalah Algoritma Decision Tree, Algoritma Support Vector Machine dan Algoritma Naïve Bayes. Data prediksi diabetes yang dikumpulkan sebanyak 2768 data dengan masing-masing algoritma memiliki 70% data training dan 30% data testing. Algoritma yang memiliki nilai evaluasi paling tinggi ialah Algoritma Naïve Bayes dengan rata-rata accuracy sebesar 78%, precision sebesar 77%, recall sebesar 78%, dan f1-score sebesar 77%.

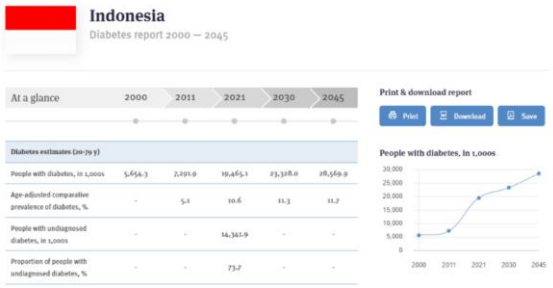
Kata kunci

diabetes, klasifikasi, Decision Tree, Naïve Bayes, Support Vector Machine

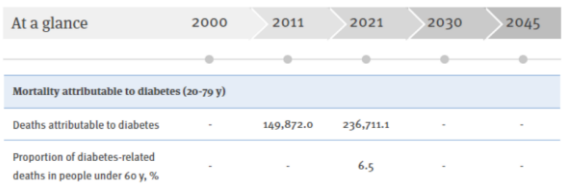
1. Pendahuluan

Diabetes adalah penyakit ketika tubuh manusia tidak dapat menggunakan insulin dengan baik. Insulin merupakan hormon yang dihasilkan dari pankreas dan berfungsi sebagai kunci untuk membawa glukosa dari makanan yang dikonsumsi oleh manusia bergerak ke sel-sel dari darah dalam tubuh, yang kemudian akan menghasilkan energi. Tubuh yang tidak dapat memproduksi insulin dapat menyebabkan kadar glukosa dalam darah menjadi tinggi. Apabila pada kasus tersebut berlangsung dalam waktu jangka panjang, maka kadar glukosa tersebut dapat merusak organ tubuh, bahkan kegagalan fungsi organ dan jaringan pada tubuh manusia yang dapat menyebabkan komplikasi bahkan kematian [1].

Menurut *International Diabetes Federation*, penderita diabetes di Indonesia pada tahun 2000 yang berusia 20-79 tahun, sebanyak 5.654,30 ribu jiwa, tahun 2011 terdapat 7.291,90 ribu jiwa, tahun 21 terdapat 19.465,10 ribu jiwa. Diperkirakan pada tahun 2030 akan bertambah banyak menjadi 23.328,00 ribu jiwa dan pada tahun 2045 akan terus mengalami kenaikan menjadi 28.569,90 ribu jiwa. Statistik penderita diabetes di Indonesia tersebut dapat dilihat pada **Gambar 1**. Menurut data tersebut, kematian yang disebabkan oleh diabetes untuk jiwa yang berusia 20-79 tahun sebanyak 149.872 ribu jiwa pada tahun 2011. Pada tahun 2021, kematian yang disebabkan oleh diabetes sebanyak 236.711 ribu jiwa yang berusia sekitar 20-79 tahun [2]. Statistik penderita diabetes yang menyebabkan kematian, dapat dilihat pada **Gambar 2**.



Gambar 1. Laporan penderita Diabetes di Indonesia



Gambar 2. Laporan penderita diabetes di Indonesia yang menyebabkan kematian

Perkembangan teknologi pada masa sekarang, dapat membantu manusia untuk mendapatkan informasi dan memprediksi penyakit tersebut. Menurut *World Health*

of Organization (WHO), *Digital Health* merupakan penggunaan teknologi digital, seluler, dan nirkabel untuk mendukung pencapaian tujuan kesehatan. Salah satu bentuk dari *Digital Health* adalah aplikasi berbasis digital dalam bidang kesehatan yang menyediakan ruang interaksi melalui media dengan koneksi internet. Penelitian ini merupakan salah satu implementasi dari *Digital Health*, yaitu menggunakan pendekatan machine learning melalui teknik klasifikasi untuk memprediksi penyakit diabetes.

Pada penelitian ini akan memprediksi penyakit diabetes dengan menggunakan Algoritma *Decision Tree*, Algoritma *Support Vector Machine* dan Algoritma *Naïve Bayes*. Dataset yang digunakan yaitu *Diabetes Prediction Dataset* yang bersumber dari *Kaggle* (https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes) berjumlah 2.768 data dengan 10 atribut dan 1 class yang menunjukkan indikasi diabetes dan tidak diabetes.

2. Metode Penelitian

Fokus pada penelitian topik tersebut adalah agar manusia mendapatkan informasi dan memprediksi penyakit diabetes awal, serta dapat membantu dalam pengembangan pengobatan dan agar mencegah terjadinya penyakit diabetes tertentu lebih dalam. Manfaat dibuatnya topik tersebut adalah pasien dapat mengelola kondisinya dan menghindari komplikasi. Untuk manusia yang tidak terkena penyakit diabetes, maka topik tersebut dapat membantu untuk mencegah terjadinya penyakit diabetes dengan menjaga kualitas hidup yang baik.

2.1. Studi Literatur

Abdi Praja, Chairisni Lubis, dan Dyah Erny Herdiwindiati menggunakan metode *Fuzzy C-Means Clustering* dan *K-Means Clustering* untuk mendeteksi penyakit diabetes [3]. Dataset yang digunakan yaitu dataset *National Institute of Diabetes and Digestive and Kidney Diseases* yang terdapat 8 indikator. Pengujian dilakukan terhadap 9 data untuk masing-masing metode. Pada metode *K-Means* memiliki hasil pengujian terbaik sebesar 73,438% dan hasil pengujian terbaik pada metode *Fuzzy C-Means* sebesar 82,812%.

Andi Maulida Argina menggunakan penerapan metode klasifikasi *K-Nearest Neighbor* dalam mengelola dataset penderita diabetes [4]. Dataset yang digunakan merupakan dataset penderita diabetes berjumlah 77 data dengan pembagian data training 90% dan data testing 10%. Hasil perhitungan pada algortima tersebut mendapatkan *accuracy* tertinggi pada K=3 sebesar 39%, *precision* tertinggi pada K=3 dan K=5 sebesar 65%, *recall* tertinggi pada K=3 sebesar 36%, dan *F-Measure* tertinggi pada K=3 sebesar 46%.

Widya Apriliah, Ilham Kurniawan, Muhammad Baydhowi, dan Tri Haryati menggunakan algoritma klasifikasi *Support Vector Machine*, *Naïve Bayes*, dan *Random Forest* untuk memprediksi kemungkinan diabetes pada tahap awal [5]. Pada penelitian tersebut menggunakan *UCI Dataset* yang berasal dari dataset dari *UCI repository* yang merupakan dataset *Diabetes Hospital in Sylhet, Bangladesh* dengan jumlah data sebanyak 520 data yang memiliki 17 atribut dan 1 kelas. Pada penelitian tersebut, algoritma klasifikasi yang menunjukkan nilai *accuracy* paling tinggi ialah algoritma *Random Forest* dengan hasil *accuracy* sebesar 97,88%, dan nilai *ROC* sebesar 0,998.

Alif Abqori Robbani, Amril Mutoi Siregar, dan Dwi Sulistya Kusumaningrum menggunakan Algoritma *C4.5* untuk Klasifikasi Penderita Penyakit Diabetes [6]. Pada penelitian tersebut, menggunakan data mengenai penderita penyakit diabetes yang 768 data dengan 8 atribut yaitu *Pregnance, Glucose, Blood P, Skin T, Insulin, BMI, DPF*, dan *Age* serta 1 label yaitu *Outcome* yang memprediksi 'Ya' dan 'Tidak'. Penelitian tersebut memiliki hasil akurasi 74,08%, laju error 26%, *precision* 43,28%, *recall* 71,16%, dan *f-measure* 53,8%.

Siti Kalimah menggunakan metode *Decision Tree* dan *Random Forest* untuk klasifikasi penyakit diabetes [7]. Pada penelitian tersebut, menggunakan algoritma *Decision Tree* dan *Random Forest* dengan data latih 80% (416 data) dan 20% data uji (104 data). Data yang digunakan berjumlah 520 data dengan 16 variabel yaitu *Age, Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Pholypagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity*, dan *Class*. Pada algoritma *Decision Tree*, hasil tingkat akurasi sebesar 91.35%, presisi sebesar 93.55%, *recall* sebesar 92.06%, specificity sebesar 90.24% dan *F1 score* sebesar 92.80%. Pada algoritma *Random Forest*, hasil tingkat akurasi sebesar 98.08%, presisi sebesar 100%, *recall* sebesar 96.88%, *specificity* sebesar 100% dan *F1 score* sebesar 98.41%.

Abu Wildan Mucholladin, Fitra Abdurrachman Bachtiar, dan Muhammad Tanzil Furqon menggunakan metode *Support Vector Machine* untuk klasifikasi penyakit diabetes [8]. Pada penelitian tersebut, data yang digunakan memiliki 768 sampel dan 9 atribut yaitu *Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age*, dan *Outcome*. Hasil pengujian menunjukkan bahwa model benchmark memiliki nilai mean *accuracy* sebesar 0,87, mean *precision* sebesar 0,82, *mean sensitivity* sebesar 0,78, dan *mean specificity* sebesar 0,92. *Model scratch* memiliki nilai mean *accuracy* sebesar 0,78, *mean precision* sebesar 0,69, *mean sensitivity* sebesar 0,59, dan *mean specificity* sebesar 0,87.

Achmad Ridwan menggunakan penerapan algoritma Naïve Bayes untuk klasifikasi penyakit diabetes mellitus [9]. Pada penelitian tersebut, memiliki data sebanyak 17 atribut dengan menggunakan algoritma *Naïve Bayes*. Hasil akurasinya sebesar 90.20% dan nilai *AUC* nya yaitu 0,95.

V. Anuja Kumari dan R. Chitra menggunakan *Support Vector Machine* untuk klasifikasi penyakit diabetes[10]. Pada penelitian tersebut, menggunakan

algoritma *Support Vector Machine* (SVM) dengan dataset diabetes. Pada algoritma *Support Vector Machine* tersebut, terdapat *Accuracy* 78%, Sensitivity 80%, dan *Specificity* 76.5%.

Mursyid Ardiansyah, Andi Sunyoto, Emha Taufiq Luthfi menggunakan Algoritma *Naïve Bayes* dan C4.5 untuk klasifikasi diabetes [11]. Pada penelitian tersebut, menggunakan algoritma *Naïve Bayes* dan C4.5. Pada algoritma *Naïve Bayes*, memiliki hasil akurasi 88.5%, *precision* 92.16%, dan *recall* 85.45%. Pada algoritma C4.5 memiliki hasil akurasi 99.03%, *precision* 100%, dan *recall* 98,18%.

Baiq Andriska Candra Permana dan Intan Komala Dewi menggunakan klasifikasi data mining *Decision Tree* dan *Naïve Bayes* untuk prediksi penyakit diabetes [12]. Pada penelitian tersebut menggunakan algoritma *Decision Tree* dan *Naïve Bayes* dengan data diabetes. Pada algoritma *Decision Tree*, hasil akurasi 95,58% dan nilai AUC 0,981. Pada algoritma *Naïve Bayes*, hasil akurasi 87,69% dan nilai AUC 0,947.

2.2. Dataset

Dataset yang digunakan merupakan *Diabetes Prediction Dataset* bersumber dari *Kaggle*, dengan jumlah 2.768 data yang memiliki 10 variabel, yang dapat dilihat pada Gambar 3 dan memiliki tipe data numerik yang dapat dilihat pada Tabel 1. *Diabetes Prediction Dataset* tersebut, dapat dilihat melalui link (https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes). Beberapa variabel tersebut adalahs sebagai berikut:

1. *Id* : Pengenal unik untuk setiap data
2. *Pregnancies* : Jumlah kehamilan
3. *Glucose* : Konsentrasi glukosa plasma selama 2 jam dalam tes toleransi glukosa oral.
4. *BloodPressure* : Tekanan darah diastolik (mmHg)
5. *SkinThickness* : Ketebalan lipatan kulit trisep (mm)
6. *Insulin* : Insulin serum 2 jam (mu U/ml)
7. *BMI* : Indeks massa tubuh (berat badan dalam kg/tinggi badan dalam m2).
8. *DiabetesPedigreeFunction* : Fungsi silsilah diabetes, skor genetik diabetes
9. *Age* : Usia pasien
10. *Outcome* : Klasifikasi biner yang menunjukkan
    (0) Tidak ada diabetes
    (1) Ada diabetes.

Tabel 1 Deskripsi Atribut Dataset

No	Atribut	Deskripsi	Satuan	Tipe
1	*Id*	*Unique identifier for each data entry*	-	*Numeric*
2	*Pregnancies*	*Number of times pregnant*	*times*	*Numeric*
3	*Glucose*	*Plasma glucose concentration over 2 hours in an oral glucose tolerance test*	*mg/dL*	*Numeric*
4	*BloodPressure*	*Diastolic blood pressure*	*mm Hg*	*Numeric*
5	*SkinThickness*	*Triceps skinfold thickness*	*mm*	*Numeric*
6	*Insulin*	*2-Hour serum insulin*	*mμ U/ml*	*Numeric*
7	*BMI*	*Body mass index*	*weight in kg / height in m^2*	*Numeric*
8	*DiabetesPedigree Function*	*Diabetes pedigree function, a genetic score of diabetes*	-	*Numeric*
9	*Age*	*Age in years*	*Years*	*Numeric*
10	*Outcome*	*Classification*	-	*Binary*

Id	Pregnanci	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesP	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1

Gambar 3. *Diabetes Prediction Dataset* 10 variabel

Data yang akan digunakan akan difokuskan pada atribut yang yang penting saja. Atribut tersebut terdiri dari 8 atribut yaitu *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, dan *Age* serta 1 label yaitu *Outcome*, yang dapat dilihat pada Gambar 3. Atribut yang tidak terlalu penting adalah Id.

Pregnanci	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesP	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1

Gambar 4. Diabetes Prediction Dataset 8 atribut dan 1 label

2.3. *Preprocessing Data*

Tujuan dilakukannya *preprocessing data* adalah agar data lebih dapat terstruktur. Pada penelitian ini, preprocessing data yang digunakan adalah memisahkan

dataset menjadi 70% *data training* dan 30% *data testing*. Nilai dari atribut tersebut bertipe *numeric*.

2.4. *Decision Tree*

Decision Tree adalah teknik pemodelan prediktif yang digunakan dalam tugas klasifikasi, pengelompokan, dan prediksi [13]. *Decision Tree* adalah pohon yang simpul akarnya dan setiap simpul internalnya diberi label dengan sebuah pertanyaan. Salah satu algoritma yang dapat digunakan untuk membuat *Decision Tree* adalah Algoritma C4.5 yang memperhitungkan atribut kategorikal dan numerik. Setiap jenis atribut, algoritma tersebut menghitung *information gain* dan *gain ratio* untuk menentukan nilai tertinggi. Algoritma tersebut terdiri dari beberapa tahapan, tahapan tersebut sebagai berikut [14] :

1. Pilih atribut atau fitur sebagai akar, didasarkan pada nilai *Gain* tertinggi
2. Buat cabang untuk setiap nilainya
3. Bagi kasus dalam cabang
4. Lakukan proses secara berulang untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

Untuk menghitung *Information Gain* yang paling baik, dapat menggunakan rumus sebagai berikut:

$$Gain(S,A) = Entropy(S) - \sum_{v \in values(A)} \frac{||S_v||}{||S||} Entropy\,(S_v) \quad (1)$$

Yang mana, *Entropy S* merupakan nilai *entropy* dari S.

$$Entropy(S) = \sum_{i} - P_i \log P_i \quad (2)$$

Keterangan:
S = Himpunan kasus
A = Atribut
n = Jumlah partisi Atribut A
$||Sv||$ = Jumlah kasus pada partisi ke–v
$||S||$ = Jumlah kasus dalam S

Untuk keseluruhan *data training* S, jenis kelas perlu diprediksi dengan persamaan *Entropy* sebagai berikut:

$$Entropy(S) = -P_{\oplus} \log P_{\oplus} - P_{\ominus} \log P_{\ominus} \quad (3)$$

Keterangan:
S = Himpunan kasus
A = Atribut
n = Jumlah partisi Atribut A
Pi = Proporsi dari Si terhadap S

2.5. *Support Vector Machine*

Support Vector Machine (SVM) merupakan klasifikasi jenis *supervised* dengan algoritma yang bekerja memecahkan masalah klasifikasi dengan mencari *maximum marginal hyperplane* (MMH) [15]. *Kernel* yang digunakan pada algoritma *Support Vector Machine* penelitian ini adalah *Kernel RBF* dan *Polynomial*.

$$f(x_d) = \sum_{i=1}^{ns} \alpha_i y_i \vec{x}_i \vec{x}_d + b \quad (4)$$

Keterangan:
ns = Jumlah support vector
αi = Nilai bobot setiap titik data
yi = Kelas data
$\vec{x}\,i$ = Variabel support vector
$\vec{x}\,d$ = Data yang akan diklasifikasikan
b = Nilai error atau bias

2.6. Naïve Bayes

Naïve Bayes merupakan metode yang memanfaatkan probabilitas dan statistik untuk menyelesaikan masalah klasifikasi [16]. Algoritma *Naïve Bayes* mengunakan teorema *Bayes* dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Algoritma tersebut melakukan klasifikasi dengan perhitungan nilai dari probabilitas P(x|y) dengan mengetahui probabilitas kelas X dan penentuannya dilakukan dengan memilih nilai max dari P(x|y) berdasarkan probabilitas [17]. Persamaan untuk menghitung nilai probabilitas tersebut dengan algoritma Naïve Bayes adalah sebagai berikut:

$$P\,(X \mid Y) = \frac{P\,(Y|X).(X)}{P(Y)} \quad (5)$$

Keterangan:
P(X|Y) = Posterior|probability yaitu nilai probabilitas X berdasarkan kondisi Y
P(Y|X) = Probabilitas Y yang ditentukan X adalah benar
P(X) = Peluang evidence penyakit
XP(Y) = Probabilitas dari nilai Y

2.7. *Confusion Matrix*

Confusion matrix disebut juga dengan *error matrix*. *Confusion matrix* yang digunakan pada penelitian ini, memiliki dua kelas yaitu kelas positif dan kelas negatif [18]. Metode evaluasi *Confusion Matrix* dengan kelas 2 dapat dilihat pada Tabel 4.

		Positive(P)	*Negative(N)*
Actual Condition	*Positive (P)*	*True Positive(TP)*	*False Negative(FN)*
	Negative (N)	*False Positive(FP)*	*True Negative(TN)*

True Positive Rate (TPR) = $\frac{TP}{P} = \frac{TP}{TP + FN}$ (6)

True Negative Rate (TNR) = $\frac{TN}{N} = \frac{TN}{TN + FP}$ (7)

$$\text{False Negative Rate (FNR)} = \frac{FN}{P} = \frac{TN}{FN + TP} \quad (8)$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{N} = \frac{FP}{FP + TN} \quad (9)$$

Keterangan:
True Positive (TP) = Jumlah dari data positif yang diklasifikasikan sebagai nilai positif
False Positive (FP) = Jumlah dari data negatif yang diklasifikasikan sebagai nilai positif
False Negative (FN) = Jumlah dari data positif yang diklasifikasikan sebagai nilai positif
True Negative (TN) = Jumlah dari data negatif yang diklasifikasikan sebagai nilai negative

$$Accuracy = \frac{TP+TN}{P+N} \quad (10)$$

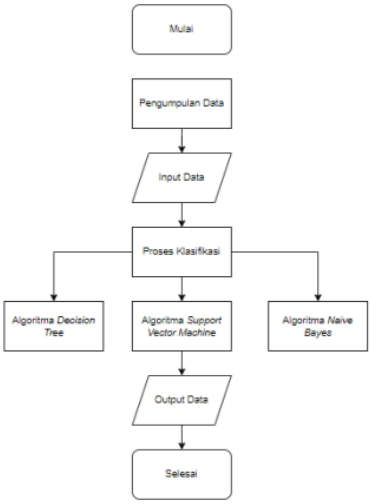
$$Precision = \frac{(TP + TN)}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$F_1\ score = 2 \times \frac{Precision\ x\ Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (13)$$

Keterangan:
Accuracy = Ukuran kinerja yang akan memberikan tingkat keakuratan dari keseluruhan model dari total jumlah data.
Precision = Ukuran kinerja yang akan memberikan informasi dari prediksi kelas postif yang sebenarnya positif.
Recall = Ukuran kinerja yang memberikan informasi dari prediksi kelas positif yang di prediksi negatif.
F1 Score = Perbandingan rata-rata presisi dan *recall* yang dibobotkan.

2.8. Diagram Prosedur Penelitian



elitian

3. Hasil Percobaan

3.1. Hasil *Decision Tree*

Dataset yang digunakan pada metode Decision Tree sebanyak 2768 data yang memiliki jumlah 10 label dan 1 class hasil diagnosa diabetes, dengan 70% data training dan 30% data testing. Algoritma yang digunakan pada metode tersebut adalah Algoritma C4. 5. Proses klasifikasi Algoritma tersebut dibangun menggunakan software *Google Colab* [19]. Evaluasi yang digunakan pada penelitian ini menggunakan *Confusion Matrix* dengan *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *F1-Score* [20]. Hasil evaluasi menunjukkan hasil *Accuracy* sebesar 69%, *Precision* sebesar 67%, *Recall* sebesar 69%, dan *F1 Score* sebesar 64% yang dapat dilihat pada Gambar 6.

	precision	recall	f1-score	support
0	0.71	0.92	0.80	614
1	0.58	0.22	0.32	300
accuracy			0.69	914
macro avg	0.64	0.57	0.56	914
weighted avg	0.67	0.69	0.64	914

Gambar 6 Hasil Evaluasi Algoritma C4.5

3.2. Hasil *Support Vector Machine*

Dataset yang digunakan pada metode *Support Vector Machine* sebanyak 2768 data yang memiliki jumlah 10 atribut dan 1 class hasil diagnosa diabetes, dengan 70% *data training* dan 30% *data testing*. Dataset yang digunakan hanya 9 atribut yaitu *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *Diabetes*, *Age*, dan *Outcome*.

Algoritma tersebut dibangun menggunakan *software Google Colab*. Klasifikasi menggunakan *Support Vector Machine* dilakukan dengan 2 kernel yaitu *RBF* dan *Polynomial*. Pengujian dilakukan sebanyak 5 kali menggunakan 3 *kernel RBF* dengan C=default, C=100, dan C=0,1 serta 2 *kernel Polynomial* dengan C=2 dan C=5. Evaluasi yang digunakan pada penelitian ini menggunakan Confusion Matrix dengan *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *F1- Score*.

Hasil evaluasi pada Kernel RBF dengan C=default, menunjukkan hasil *Accuracy* sebesar 76%, *Precision* sebesar 75%, *Recall* sebesar 76%, dan F1 Score sebesar 75% yang dapat dilihat pada Gambar 7.

	precision	recall	f1-score	support
0	0.78	0.87	0.82	535
1	0.71	0.55	0.62	296
accuracy			0.76	831
macro avg	0.74	0.71	0.72	831
weighted avg	0.75	0.76	0.75	831

Gambar 7 Hasil Evaluasi *Kernel RBF* dengan C=*default*

Hasil evaluasi pada Kernel RBF dengan C=100, menunjukkan hasil *Accuracy* sebesar 77%, *Precision* sebesar 77%, *Recall* sebesar 77%, dan F1 Score sebesar 76% yang dapat dilihat pada Gambar 8.

	precision	recall	f1-score	support
0	0.79	0.89	0.83	535
1	0.73	0.57	0.64	296
accuracy			0.77	831
macro avg	0.76	0.73	0.74	831
weighted avg	0.77	0.77	0.76	831

Gambar 8 Hasil Evaluasi *Kernel RBF* dengan C=100

Hasil evaluasi pada Kernel RBF dengan C=0.1, menunjukkan hasil *Accuracy* sebesar 74%, *Precision* sebesar 73%, *Recall* sebesar 74%, dan *F1 Score* sebesar 72% yang dapat dilihat pada Gambar 9.

	precision	recall	f1-score	support
0	0.74	0.91	0.82	535
1	0.72	0.43	0.54	296
accuracy			0.74	831
macro avg	0.73	0.67	0.68	831
weighted avg	0.73	0.74	0.72	831

Gambar 9 Hasil Evaluasi Kernel RBF dengan C=0.1

Hasil evaluasi pada Kernel Polynomial dengan C=2, menunjukkan hasil *Accuracy* sebesar 76%, *Precision* sebesar 75%, *Recall* sebesar 76%, dan F1 Score sebesar 74% yang dapat dilihat pada Gambar 10.

	precision	recall	f1-score	support
0	0.77	0.89	0.82	535
1	0.72	0.52	0.60	296
accuracy			0.76	831
macro avg	0.74	0.70	0.71	831
weighted avg	0.75	0.76	0.74	831

Gambar 10 Hasil Evaluasi Kernel Polynomial dengan C=2

Hasil evaluasi pada Kernel Polynomial dengan C=2, menunjukkan hasil *Accuracy* sebesar 77%, *Precision* sebesar 77%, *Recall* sebesar 77%, dan F1 Score sebesar 75% yang dapat dilihat pada Gambar 11.

	precision	recall	f1-score	support
0	0.76	0.93	0.84	535
1	0.78	0.47	0.59	296
accuracy			0.77	831
macro avg	0.77	0.70	0.71	831
weighted avg	0.77	0.77	0.75	831

Gambar 11 Hasil Evaluasi Kernel Polynomial dengan C=5

Berdasarkan hasil evaluasi, didapatkan nilai tertinggi pada Kernel RBF dengan C=100 yang mendapatkan nilai *Accuracy* sebesar 77%, *Precision* sebesar 77%, *Recall* sebesar 77%, dan F1- Score sebesar 76%. Nilai evaluasi terendah ditunjukkan pada Kernel RBF dengan C=0.1 yang mendapatkan nilai *Accuracy* sebesar 74%, *Precision* sebesar 73%, *Recall* sebesar 74%, dan F1-Score sebesar 72%.

3.3. Hasil *Naïve Bayes*

Dataset yang digunakan pada metode *Naïve Bayes* sebanyak 2768 data yang memiliki jumlah 10 atribut dan 1 class hasil diagnosa diabetes, dengan 70% *data training* dan 30% *data testing*. Dataset yang digunakan hanya 9 atribut yaitu *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *Diabetes*, *Age*, dan *Outcome*. Algoritma tersebut dibangun menggunakan *software Google Colab*. Evaluasi yang digunakan pada penelitian ini menggunakan *Confusion Matrix dengan Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *F1-Score*. Hasil evaluasi menggunakan algoritma tersebut, menunjukkan hasil nilai *Accuracy* sebesar 78%, *Precision* sebesar 77%, *Recall* sebesar 78%, dan F1-Score sebesar 77%, yang dapat dilihat pada Gambar 12.

	precision	recall	f1-score	support
0	0.82	0.86	0.84	562
1	0.68	0.60	0.64	269
accuracy			0.78	831
macro avg	0.75	0.73	0.74	831
weighted avg	0.77	0.78	0.77	831

Gambar 12 Hasil Evaluasi *Naive Bayes*

4. Kesimpulan

Berdasarkan hasil dan pembahasan pada Diabetes Prediction Dataset yang bersumber dari Kaggle (https://www.kaggle.com/datasets/nanditapore/healthcare-diabetes) dengan dataset berjumlah 2768 data, menunjukkan hasil evaluasi yang dihasilkan dari masing-masing metode yaitu:

1. Memiliki nilai hasil evaluasi tertinggi pada metode *Naïve Bayes* diantara metode *Decision Tree* dan *Support Vector Machine* dengan nilai *Accuracy* sebesar 78%, *Precision* sebesar 77%, *Recall* sebesar 78%, dan *F1-Score* sebesar 77%.
2. Kesimpulannya, algoritma yang paling tepat untuk dataset tersebut ialah algoritma metode Naïve Bayes.

DAFTAR PUSTAKA

[1] F. Fatmawati, "PERBANDINGAN ALGORITMA KLASIFIKASI DATA MINING MODEL C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT DIABETES," Jurnal Techno Nusa Mandiri, vol. 13, no. 1, pp. 50-59, 2016.

[2] "Indonesia diabetes report 2000 - 2045," International Diabetes Federation, [Online]. Available: https://diabetesatlas.org/data/en/country/94/id.html. [Accessed 22 September 2023].

[3] A. Praja, C. Lubis and E. D. Herdiwindiati, "Deteksi Penyakit Diabetes dengan Metode Fuzzy C Means Clustering dan K-Means Clustering," Computatio: Journal of Computer

Science and Information Systems, vol. 1, no. 1, pp. 15-24, 2017.
[4] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," Indonesian Journal of Data and Science, vol. 1, no. 2, pp. 29-33, 2020.
[5] W. Apriliah, I. Kurniawan and M. Baydhowi, "Apriliah, W., Kurniawan, I., BaydhoPrediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi Random Forest," Sistemasi: Jurnal Sistem Informasi, vol. 10, no. 1, pp. 163-171, 2021.
[6] A. A. Robbani, A. M. Siregar and D. S. Kusumaningrum, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5," Scientific Student Journal for Information, vol. 3, no. 1, pp. 76-82, 2022.
[7] S. Kalimah, "Klasifikasi Penyakit Diabetes Menggunakan Metode Decision Tree dan Random Forest," Repository Universitas Sriwijaya, 2022.
[8] A. W. Mucholladin, F. A. Bachtiar and M. T. Furqon, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 5, no. 2, pp. 622-633, 2021.
[9] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," Jurnal Sistem Komputer dan Kecerdasan Buatan (SISKOM-KB), vol. 4, no. 1, 2020.
[10] V. A. Kumari and R. Chitra, "Classification Of Diabetes Disease Using Support Vector Machine," International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797-1801, 2013.
[11] M. Ardiansyah, A. Sunyoto and E. T. Luthfi, "Analisis Perbandingan Akurasi Algoritma Naïve Bayes dan C4.5 untuk Klasifikasi Diabetes," Edumatic: Jurnal Pendidikan Informatika, vol. 5, no. 2, pp. 147-156, 2021.
[12] B. A. C. Permana and I. K. Dewi, "Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naive Bayes Untuk Prediksi Penyakit Diabetes," Infotek J. Inform. dan Teknol, vol. 4, no. 1, pp. 63- 69, 2021.
[13] E. E. Barito, J. T. Beng and D. Arisandi, "Penerapan Algoritma C4. 5 untuk Klasifikasi Mahasiswa Penerima Bantuan Sosial Covid-19," Jurnal Ilmu Komputer dan Sistem Informasi, vol. 10, no. 1, 2022.
[14] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4. 5," Jurnal Edik Informatika Penelitian Bidang Komputer Sains dan Pendidikan Informatika, vol. 2, no. 2, pp. 213-219, 2017.
[15] A. J. Hendryli and D. E. Herwindiati, "Klasifikasi Tanaman Obat Herbal Menggunakan Metode Support Vector Machine," Computatio: Journal of Computer Science and Information Systems, vol. 5, no. 1, pp. 25-35, 2021.
[16] I. Andriyanto, E. Santoso and S. , "Pemodelan Sistem Pakar Untuk Menentukan Penyakit Diabetes Mellitus Menggunakan Metode Naive Bayes Studi Kasus: Puskesmas Poncokusumo Malang," Andriyanto, I., Santoso, E. and Suprapto, S., 2018. Pemodelan Sistem Pakar Untuk Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 2, pp. 880-887, 2018.
[17] H. Apriyani and K. Kurniati, "Perbandingan Metode Naive Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," Journal of Information Technology Ampera, vol. 1, no. 3, pp. 133-143, 2020.
[18] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," Normawati, Dwi, and Surya Allit Prayogi. "Implementasi Naïve Bayes Classifier Dan J-SAKTI (Jurnal Sains Komputer Dan Informatika, vol. 5, no. 2, pp. 697-711, 2021.
[19] A. S. Rahayu, A. Fauzi and R. , "Rahayu, Ayu Sri, AhmaKomparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify," Jurnal Sistem Komputer dan Informatika (JSON), vol. 4, no. 2, pp. 349-354, 2022.
[20] T. A. Y. Siswa and N. A. Verdikha, "Komparasi Algoritma Klasifikasi Untuk Menentukan Evaluasi Kinerja Terbaik Pada Status Akreditasi Sekolah/Madrasah Kalimantan Timur Berdasarkan IASP 2020," Jurnal Informatika Teknologi dan Sains (Jinteks), vol. 4, no. 3, pp. 185-192, 2022

Tasya Syamsudin, saat ini sebagai mahasiswa Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara.

Teny Handhayani, memperoleh gelar S.Kom. dari Institut Pertanian Bogor tahun 2008. Kemudian memperoleh gelar Magister Ilmu Komputer dari Universitas Indonesia tahun 2013, dan kemudian memperoleh gelar Philsophy, Ilmu Komputer dari University of York tahun 2021. Saat ini aktif sebagai Dosen Tetap di Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta.

Muhammad Isnaini Syaifudin, saat ini sebagai mahasiswa Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara.