P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

IMPELEMENTASI K-NEAREST NEIGHBORS, DECISION TREE DAN SUPPORT VECTOR MECHINE PADA DATA DIABETES

Miftahul Irfan[1], Wardhani Utami Dewi[2*], Khoirin Nisa[3], Mustofa Usman[4].
[1, 2*, 3, 4)] Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung

[1,2,3,4)] Jl Prof Dr Sumantri Brojonegoro No 1, Gedong Meneng, Rajabasa, Bandar Lampung
[1] E-Mail : miftahulirfan311@gmail.com , [2*] E-mail: dewiutamiwardhani@gmail.com

Abstrak *:* Diabetes merupakan salah satu penyakit yang menjadi penyebab kematian terbesar didunia. Kasus kematiannya pun tercatat lebih dari 4 juta pada tahun 2019. Diabetes juga dapat menyebabkan timbulnya penyakit lainnya. Bahaya diabetes ini menjadi perhatian khusus WHO. Seiring dengan perkembangan teknologi ini, banyak sekali kolaborasi antara bidang kesehatan, statistic dan computer untuk menanggulangi berbagai macam penyakit. Algortima machine learning menjadi popular dalam proses klasifikasi data dan sudah banyak diterapkan pada data kesehatan. Dengan begitu pada artikel ini akan dilakukan perbandingan algoritma machine learning KNN, Decision Tree, dan SVM untuk melihat algortima mana yang paling cocok untuk klasifikasi data diabetes. Hasil menunjukkan bahwa KNN dan SVM memiliki akurasi yang cukup besar yaitu 81,13%. Sehingga kedua algortima tersebut dapat menjadi rekomendasi proses klasifikasi data diabetes sehingga dapat membantu dokter dalam menanggulangi penyakit diabetes. Hasil ini juga menunjukkan bahwa 8 variabel yang digunakan berpengaruh terhadap resiko diabetes.

Kata Kunci :KNN, SVM, Machine Learning, Klasfikasic

Abstract**:** *Diabetes is one of the leading causes of death in the world. Cases of death were recorded at more than 4 million in 2019. Diabetes can also cause other diseases. The dangers of diabetes are of particular concern to the WHO. Along with the development of this technology, there are lots of collaborations between the health sector, statistics and computers to overcome various diseases. Machine learning algorithms are becoming popular in data classification processes and have been widely applied to health data. With this in mind, this article will compare the KNN, Decision Tree, and SVM algorithms to see which algorithm is most suitable for classifying diabetes data. The results show that KNN and SVM have a fairly large accuracy, namely 81.13%. So that the two algorithms can become recommendations for the process of classifying diabetes data so that they can help doctors in tackling diabetes. These results also indicate that the 8 variables used influence the risk of diabetes.*

Keywords: *KNN, SVM, Machine Learning, Classification.*

PENDAHULUAN

Diabetes merupakan salah satu penyakit yang banyak diderita dan cukup berbahaya. Bahkan diabetes dapat menimbulkan penyakit-penyakit lainnya seperti gagal ginjal. Diabetes adalah kondisi dimana metabolism kronis yang ditandai dengan meningkatnya kadar glukosa darah dalam diri seseorang (Campbell et al., 2020). Diabetes merupakan salah satu penyakit yang termasuk kedalam 10 penyakit teratas

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

yang menyebabkan kematian. Tercatat bahwa pada tahum 2019 sebanyak 4.2 juta jiwa meninggal karena diabetes (Tanase et al., 2020). Menurut Hussein terdapat beberapa factor yang mempengaruhi diabetes dimana hasil menunjukkan bahwa BMI menjadi factor yang sangat berpengaruh (Hussein et al., 2022). Studi tentang diabetes banyak dilakukan dengan metode machine learning. Studi itu pernah dilakukan di universitas Toronto dimana mereka menunjukkan bahwa algorima dalam machine learning dapat memperkirakan dengan tepat berpa banyak pasien yang terkena diabetes. Algoritma machine learning juga dapat melihat apakah seseorang memiliki resiko terhadap tinggi atau rendah tertular penyakit dengan memasukkan variabel factornya (Sekaran & Bougie, 2016). Dengan system klasifikasi ini dapat membantu dokter dalam menganalisi data penyakit tersebut (Edla & Cheruku, 2017).

Seiring dengan perkembangan zaman dan kemajuan teknologi yang terjadi pada saat ini. Banyak metode-metode dalam ilmu pengetahuan yang terus dikembangkan untuk melakukan analisis data. Metode statistic klasik umumnya memiliki asumsi-asumsi yang sulit untuk dipenuhi. Tidak hanya itu, di era big data saat ini menyulitkan metode statistic klasik dalam melakukan pekerjaanya. Oleh sebab itu muncul metode baru yang disebut machine learning. Machine learning dianggap sebagai salah satu pendekatan kontemporer dalam memprediksi, mengidentifikasi, dan mengambil keputusan tanpa keterlibatan manusia. Machine learning berkembang pesat dalam industri medis mulai dari diagnosis hingga visualisasi penyakit dan studi tentang penularan penyakit (Battineni et al., 2019). Dalam mesin learning ada banyak metode yang bisa digunakan diantaranya k-Nearest Neighbors (KNN), Decision Tree dan Support Vector Mechine (SVM).

Algoritma KNN merupakan salah satu metode yang berguna untuk klasifikasi. Algoritma KNN juga merupakan salah satu metode klasifikasi yang paling terkenal. Prinsip dalam KNN adalah menyusun data dalam ruang yang ditentukan dengan memilih fitur (Chanal et al., 2022). Algoritma KNN dapat secara efektif melindungi privasi data, namun kinerjanya memiliki degradasi yang tidak bisa diabaikan (Zhao et al., 2021). Algoritma KNN sendiri memiliki fitur-fitur yang membuatnya menjadi algoritma yang kuat jika dibandingkan dengan metode klasifikasi lainnya. Kelebihan dari algoritma KNN sendiri adalah kesederhanaan, kelengkapan dan ketahanan terhadap data noise, serta memberikan kinerja yang relative tinggi (Wu et al., 2017). Namun KNN juga memiliki kekurangan, diantarnya waktu eksekusi yang relative lambat, sensitive terhadap data yang cukup besar, sensitive terhadap nilai k, dan langkah komputasi yang cukup tinggi. Banyak sekali penelitian yang menunjukkan bahwa KNN merupakan alat klasifikasi yang baik. (Freitas et al., 2022), menyatakan bahwa algoritma KNN merupakan alat yang canggih dalam mengklasifikasikan dan memilih film. (Gómez-Meire et al., 2014) dan (Y. Chen et al., 2017) menunjukkan bahwa algoritma KNN adalah alat yang manjanjikan dalam penelitian makanan pemilihan wine dan cuka. Dalam bidang kesehatan pun banyak digunakan KNN. Studi tentang algoritma KNN dilakukan untuk mengklasifikasikan kanker payudara (Medjahed et al, 2013), deteksi tumor otak (Deepa et al., 2022), sedangkan (Tsalera et al., 2020), menggunakan algoritma KNN untuk klasifikasi kebisingan lingkungan perkotaan.

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

Selain itu, pengklasifikasian data juga dapat menggunakan decision tree. Decision tree menggunakan konsep pohon dalam algoritmanya dimana daun-daunnya sebagai keputusan yang diperoleh berdasarkan serangkaian aturan-aturan yang dilewati pada bagain batang dan daun. Decision tree memiliki manfaat sebagai model yang dapat dipahami dan mudah ditafsirkan. Hasil dari decision tree pun mudah untuk dijelaskan kepada khalayak umum dan praktisi (Kappelhof et al., 2021). Decision tree juga banyak digunakan dalam bidang medis. (Lowie et al., 2021), menyatakan bahwa dengan menggunakan decision tree dapat menghemat waktu dan biaya terutama pada penelitian bidang kesehatan yang mengggunakan sampel data. Decisoin tree cukup baik dalam memprediksi kelangsungan kidup keseluruhan pasien dengan limfoma sel b besar difus di sistem saraf pusat (Huang et al., 2022), untuk memprediksi risiko perdarahan intrakranial setelah cedera otak traumatis ringan (Turcato et al., 2021), Algoritme Decision tree untuk memprediksi kematian pada pasien COVID-19 (Elhazmi et al., 2022) serta untuk memprediksi gangguan bipolar dari gangguan depresi mayor (Zhu et al., 2022).

Kemudian metode lainnya yang cukup banyak digunakan dalam pengklasifikasian data dan masalah regresi adalah SVM. SVM telah terbukti menjadi alat yang sangat efektif untuk klasifikasi terawasi (Alcaraz et al., 2022). SVM sendiri merupakan suatu metode yang memisihkan data dalam ruang multidemensi. Pemisahan data dengan SVM ini melibatkan garis yang disebut hyperplane. Semakin maksimum hyperplane yang diperoleh maka dapat memudahkan pemisahan data dimasa depan. SVM juga melibatkan beberapa fungsi kernel dalam pengklasifikasian data yang tidak dapat dipisahkan secara linear. Pada tahun 2011 Chao Chen menggunakan SVM dalam bidang medis untuk prediksi pencarian pada fraksi aktif dan resep bahan obat herbal (C. Chen et al., 2011), Model SVM dalam mendiagnosis bronkiektasis dengan infeksi saluran napas kronis (Hu et al., 2023), SVM untuk otentikasi obat-obatan herbal (J. Zhang et al., 2022), SVM juga dapat diterapkan untuk prediksi demensia (Battineni et al., 2019).

Ketiga metode tersebut terkenal sebagai metode yang baik dalam proses klasifikasi data. Beberapa penelitian sudah melakukan perbandingan metode-metode tersebut. (Adem, 2020) telah membandingkan metode SVM, SAE dan kNN pada data kanker payudara dimana SAE dan kNN memiliki akurasi yang lebih baik dibanding dengan SVM. (Bansal et al., 2022), membandingkan KNN, Genetic, SVM, Decision Tree, and Long Short Term Memory. Hasil yang diperoleh menunjukkan bahwa jaringan LSTM dan algoritme SVM telah memberikan salah satu hasil terbaik dalam hal analitik prediktif dalam aplikasi waktu. Sehingga pada penelitan ini juga akan dilakukan perbandingan ketiga metode tersebut pada kasus diabetes. Hasil akurasi yang besar menunjukkan bahwa metode ini merupakan metode yang baik dan juga mengindikasikan bahwa variabel-variabel yang digunakan berpengaruh terhadap identifikasi pasien diabetes.

KAJIAN PUSTAKA DAN LANDASAN TEORI

Algoritma KNN

Algoritma KNN merupakan teknik klasifikasi data yang tidak berlabel dengan menghitung jarak antara setiap titik data yang tidak berlabel dan semua titik lainnya dalam kumpulan data (Almomany et al.,

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

2022). Bagian terpenting dalam algoritma KNN adalah dengan penentuan K optimum. K merupakan jumlah tetangga dari titik data baru. Nilai K yang sangat rendah dapat menyebabkan data noice. Sedangkan nilai K yang sangat tinggi akan menimbulkan kebingungan (Bansal et al., 2022). Untuk menghitung jarak dalam KNN biasanya menggunakan jarak Euclidean. Klasifikasi dilakukan berdasarkan k-neighbor terdekat (jarak terkecil). Jarak antara titik sample test x dengan titik training X dimana $x_i = (x_1, x_2, \dots x_n)$ dan $X_i = (X_1, X_2, \dots X_n)$ dapat dihitung dengan persamaan berikut, (S. Zhang et al., 2017) yaitu:

$$d(x, X) = \sqrt{\sum_{i=1}^{n} (x_i - X_i)^2}$$

Algoritma Decision Tree

Algoritma Decision tree adalah suatu teknik data mining yang dapat melakukan visualisasi hubungan antara banyak variabel dengan serangkaian aturan eksplisit untuk klasifikasi data pengaruhnya terhadap variabel dependen (Song & Lu, 2015). Hasil dari algoritma ini menghasilkan diagram hierarkis yang mudah diinterpretasikan dan dapat memperjelas hubungan yang kompleks antara variabelnya (Shirali et al., 2018). Ada banyak implementasi Decision Tree, tetapi salah satu yang paling terkenal adalah algoritma C5.0. Algoritma ini dikembangkan oleh ilmuwan computer J. Ross Quinlan sebagai versi perbaikan dari algoritma sebelumnya, C4.5, yang merupakan peningkatan dari algoritma ID3 (Iterative Dichotomiser 3) *miliknya.*

C5.0 menggunakan entropi untuk mengukur kemurnian. Entropi sampel data menunjukkan bagaimana campuran nilai kelas; nilai minimum 0 menunjukkan bahwa sampel benar-benar homogen, sedangkan 1 menunjukkan jumlah gangguan maksimum (Lantz, 2013). Definisi entropi ditentukan oleh:

$$Entropy(s) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

information gain untuk sebuah fitur F dihitung sebagai perbedaan antara entropi pada segmen sebelum pemisahan (S1),dan partisi yang dihasilkan dari split (S2):

$$InfoGain(F) = Entropy(S_1) - Entropy(S_2)$$

Algoritma SVM

SVM merupakan pemisahan data dalam ruang multidimensi. Tugas dari algoritma SVM adalah mengidentifikasi garis yang memisahkan kedua kelas. Dalam proses pemisahannya mengandalkan hyperplane. Hyperplane dan support vector dalam algoritma SVM merupakan bagian terpenting dalam proses ini. Support vector adalah titik-titik dari setiap kelas yang paling dekat dengan MMH. Karena sifatnya yang mendukung terhadap hyperplane, mereka disebut sebagai vektor pendukung. Menurut Lantz (2013), definisi hyperplan dimensi-n adalah sebegai berikut:

$$\vec{w}.\vec{x} + b = 0$$

dimana w adalah vector n bobot yaitu $\{w1, w2, \dots, wn\}$ dan b adalah bias. Persamaan diatas dapat ditulis kembali menjadi:

$$\vec{w}.\vec{x} + b \geq +1$$
$$\vec{w}.\vec{x} + b < -1$$

hyperplane ini ditentukan sedemikian rupa sehingga semua titik dari satu kelas berada di atas hyperplane pertama dan semua titik dari kelas lain berada di bawah hyperplane kedua. Oleh karena itu, untuk memaksimalkan jarak, kita perlu meminimalkan $||w||$ yaitu:

$$\min \frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{n} \xi_i$$

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

$$s.t.\, y_i(\overrightarrow{w}.\overrightarrow{x_i} - b) \geq 1 - \xi_i\,, \forall \overrightarrow{x_i}\,, \xi_i \geq 0$$

Dalam SVM terdapat fungsi kernel sebagai alternative jika data tidak dapat dipisahkan secara linear. Berikut adalah fungsi kernelnya:

1. Kernel Linear (tanpa transformasi data).
$$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i.\vec{x}_j$$
2. Kernel Polinomial (derajat d menambahkan transformasi data non-linear sederhana).
$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i.\vec{x}_j + 1)^d$$
3. Kernel sigmoid (menghasilkan model SVM yang agak analog dengan jaringan saraf menggunakan fungsi aktivasi sigmoid).
$$K(\vec{x}_i, \vec{x}_j) = \tanh(k\vec{x}_i.\vec{x}_j - \delta)$$
4. Kernel RBF Gaussian (Kernel RBF berkinerja baik pada banyak jenis data dan dianggap sebagai titik awal yang masuk akal untuk banyak tugas pembelajaran).
$$K(\vec{x}_i, \vec{x}_j) = e^{\frac{-\left|\left|\vec{x}_i.\vec{x}_j\right|\right|^2}{2\sigma^2}}$$

METODE

Data yang digunakan dalam penelitian ini adalah data diabetes yang diperoleh dari website www.kaggle.com/datasets/whenamancodes/predict-diabities. Data ini terdiri dari 768 sampel, 1 varibel terikat yaitu menderita diabetes (Yes) atau tidak menderita diabetes (No), dan 8 variabel bebas yang terdiri dari Pregnancies (X_1), Glucose (X_2), Blood Pressure (X_3), Skin Thickness (X_4), Insulin (X_5), BMI (X_6), Diabetes Pedigree Function (X_7), and Age (X_8).

Metode yang digunakan dalam artikel ini adalah algoritma kNN, Decision Tree dan SVM. Pertama dilakukan pembagian data menjadi data training dan data testing. Data training sebanyak 450 sedangkan data testing sebanyak 318. Kemudian akan dibandingan ketiga metode. Metode yang terbaik memiliki akurasi yang paling besar didukung dengan sensitivity dan specificity yang juga besar

HASIL DAN PEMBAHASAN

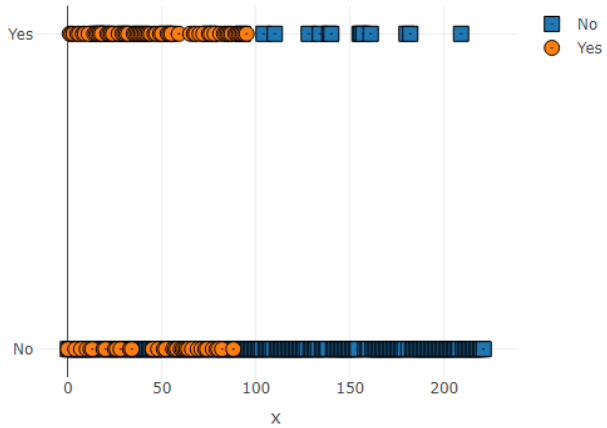
Algoritma K-Nearest Neighbors (KNN)

Dalam proses klasifikasi dengan algoritma KNN terlebih dahulu akan ditentukan jumlah k optimum (k terbaik) yang akan digunakan. Dalam penelitian ini digunakan beberapa k seperti yang ditunjukkan dalam tabel berikut:

Tabel 1. Akurasi Model

Jumlah k	Akurasi
1	73.58
5	77.36
7	76.42
15	81.13
27	81.13

Dari jumlah k yang digunakan diatas, dapat dilihat bahwa jumlah k optimum adalah 15 dan 27 dengan tingkat akurasi yang sama besar yaitu sebesar 81.13. Namun pada penelitian ini menggunakan jumlah k yang lebih kecil yaitu 15. Sehingga yang akan digunakan dalam klasifikasi data diabetes ini adalah algoritma KNN dengan k=15. Gambar berikut ini menunjukkan klasifikasi dengan KNN dengan k=15.



Gambar 1. Klasifikasi dengan kNN

Gambar diatas menunjukkan bahwa terdapat kesalahan dalam klasifikasi

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

dengan KNN. Dalam hal ini kNN tidak dapat memisahkan data dengan tepat 100%. Dapat dilihat bahwa terdapat data yang diklasifikasi kedalam kelompok menderita diabetes namun pada data aslinya dia tidak menderita diabetes. Begitu juga sebaliknya, terdapat data yang diklasifikasikan kedalam kelompok yang tidak menderita diabetes namun pada data aslinya termasuk kelompok menderita diabetes. Namun kelemahan dari gambar diatas adalah kita tidak dapat melihat pengklasifikasian data diabetes per variabel. Hasil klasifikasi data diabetes dengan algoritma kNN menunjukkan hasil akurasi yang cukup bagus. Hal ini mengindikasikan bahwa variabel-variabel yang digunakan mempunyai pengaruh terhadap penderita diabetes. Dimana terdapat kondisi-kondisi yang apabila hal tersebut dialami seseorang maka seseorang beresiko terkena diabetes. Untuk melihat kesalahan klasifikasi dalam KNN dapat dilihat dengan menggunakan cross table pada data testing seperti berikut ini:

Tabel 2. Cross Table KNN

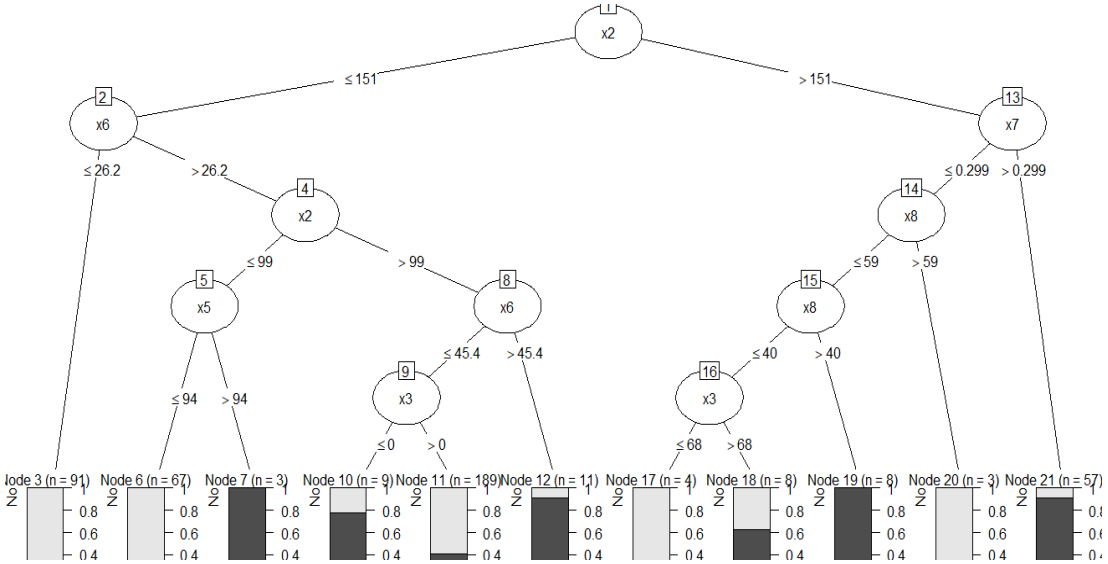
Data Testing	Klasifikasi KNN (k=15) No	Yes	Row Total
No	197	25	222
	0.887	0.113	0.698
	0.849	0.291	
	0.619	0.079	
Yes	35	61	96
	0.365	0.635	0.302
	0.151	0.709	
	0.11	0.192	
Column Total	232	86	318
	0.730	0.270	

Dari tabel diatas, dapat dilihat bahwa 197 orang yang tidak menderita diabetes berdasarkan data testing di klasifikasikan dengan tepat bahwa mereka tidak menderita diabetes. Namun terdapat 25 kesalahan klasifikasi, dimana pada kenyataanya termasuk dalam kelompok tidak menderita diabetes namun diklasifikasikan kedalam kelompok penderita diabetes. Selanjutnya dapat dilihat pula bahwa 61 orang yang menderita diabetes tepat di klasifikasikan kedalam kelompok orang yang menderita diabetes. Namun terdapat 35 kesalahan klasifikasi, dimana seharusnya 35 orang tersebut termasuk kelompok yang menderita diabetes namun diklasifikasikan dalam kelompok yang tidak menderita diabetes.

Decision Tree

Dalam proses klasifikasi dengan metode decision tree ini tidak melibatkan semua variabel. Variabel yang tidak terlibat tersebut dianggap homogen sehingga tidak perlu dipisahkan. Gambar berikut ini menunjukkan pohon klasifikasi dalam decision tree untuk factor-faktor yang mempengaruhi diabetes:

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150



* Pregnancies (X1), Glucose (X2), Blood Pressure (X3), Skin Thickness (X4), Insulin (X5), BMI (X6), Diabetes Pedigree Function (X7), and Age (X8).

Gambar 2. Klasifikasi dengan Decision Tree

Dari gambar diatas, kita dapat melihat bahwa tidak ada variabel Pregnanceis (X1) dan Skin Thickness (X4). Kedua variabel tersebut dianggap homogen sehingga tidak diklasifikasikan dalam diagram decision tree. Sedangkan untuk ke-enam variable lainnya berpengaruh terhadap penentuan kasus diabetes dalam data ini. Dari gambar diatas jelas bahwa dari 91 orang yang memiliki kadar Glucose ≤ 151 dan BMI ≤ 26.2 terdapat 87 orang yang tidak terkena diabetes sedangkan 4 orang lainnya terkena diabetes. Dengan begitu kondisi ini merupakan kondisi yang baik untuk terhindar dari diabetes. Namun tidak serta demikian. Terlihat bahwa meskipun BMI > 26.2 namun memiliki Glucose ≤ 99 dan Insulin ≤ 94 termasuk kedalam kelompok yang tidak menderita diabetes dimana terdiri dari 61 orang yang tidak menderita diabetes berbanding 6 orang yang menderita diabetes. Berbanding terbalik jika BMI > 26.2 dan Glucose ≤ 99 namun Insulin > 94 maka tergolong kelompok yang menderita diabetes. Kemudian kita perhatikan lagi untuk seseorang yang memiliki Glucose $99 < x \leq 151$ dan memiliki BMI > 45.4 tergolong kelompok penderita diabetes dengan perbandingan 10 penderita diabetes dan 1 tidak menderita diabetes. Pada kondisi Glucose yang sama dengan BMI antara 26.2 sampai 25.4, kasus diabetes tergantung pada Blood Pressure. Jika Blood Pressure ≤ 0 maka termasuk kelompok yang mayoritas menderita diabetes dengan perbandingan 7 yang menderita dan 2 tidak menderita diabetes. Namun selainnya jika Blood Pressure pada kondisi ini > 0 maka termasuk kedalam kelompok tidak menderita diabetes. Namun perbandingan pada kelompok ini tidak cukup besar dimana dari 189 orang terdapat 112 yang tidak menderita diabetes dan 77 lainnya menderita diabtes.

Disisi yang berbeda dimana kondisi Glucose > 151 dan memiliki Diabetes Pedigree Function > 0.299 akan termasuk kelompok yang menderita diabetes dengan jumlah penderita 52 berbanding 5 yang tidak menderita. Namun di kondisi Glucose yang sama namun memiliki Diabetes Pedigree Function ≤ 0.299 maka akan

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

tergantung pada Usia. Terlihat bahwa seseorang yang Usia > 59 tahun termasuk kelompok yang tidak menderita diabetes. Namun yang berusia ≤ 59 tahun akan dibagi kedalam dua kelompok lagi yaitu ≤ 40 tahun dan $40 < x_8 \leq 59$ tahun. Untuk seseorang yang berusia $40 < x_8 \leq 59$ dengan kondisi Glucose > 151 dan Diabetes Pedigree Function ≤ 0.299 akan tergolong ke dalam kelompok orang yang menderita diabetes dengan jumlah kasus 8 orang dari 8. Sedangkan untuk kelompok usia yang kurang dari atau sama dengan 40 tahun akan ditentukan oleh Blood Pressure. Dimana dengan kondisi yang sama namun memiliki Blood Pressure ≤ 68 termasuk kedalam kelompok yang tidak menderita diabetes. Namun disisi berlawanan dimana Blood Pressure > 68 maka akan termasuk kedalam kelompok yang menderita diabetes. Perhatikan tabel berikut ini untuk melihat kesalahan klasifikasi dengan menggunakan Decision Tree.

Tabel 3. Cross Table Decision Tree

Data Testing	Klasifikasi Decision Tree No	Yes	Row Total
No	191	31	222
	0.86	0.14	0.698
	0.776	0.431	
	0.601	0.097	
Yes	55	41	96
	0.573	0.427	0.302
	0.224	0.569	
	0.173	0.129	
Column Total	246	72	318
	0.774	0.226	

Dari tabel diatas, dapat dilihat bahwa 191 orang yang tidak menderita diabetes tepat diklasifikasikan kedalam kelompok orang yang tidak menderita diabetes. Sedangkan tepat disebelah kanannya terlihat bahwa terdapat 31 orang yang tidak menderita diabetes namun diklasifikasikan kedalam kelompok penderita diabetes. Ini menunjukkan kesalahan klasifikasi yang terjadi. Tingkat kesalahan klasifikasinya tidak besar jika dibanding dengan data yang diklasifikasikan dengan benar. Selanjutnya dapat dilihat pula bahwa terdapat 41 orang yang menderita diabetes dan benar diklasifikasikan kedalam kelompok yang menderita diabetes. Namun hal yang menarik terjadi pada kesalahan klasifikasinya dimana jumlahnya lebih banyak dibanding yang diklasifikasikan dengan tepat yaitu sebanyak 55 orang yang seharusnya termasuk kedalam kelompok orang yang menderita diabetes namun diklasifikasikan kedalam kelompok yang tidak menderita diabetes. Secara keseluruhan, proses klasifikasi dengan decision tree ini terbilang cukup baik namun tidak lebih baik dibanding dengan algoritma kNN.

Suppport Vector Mechine

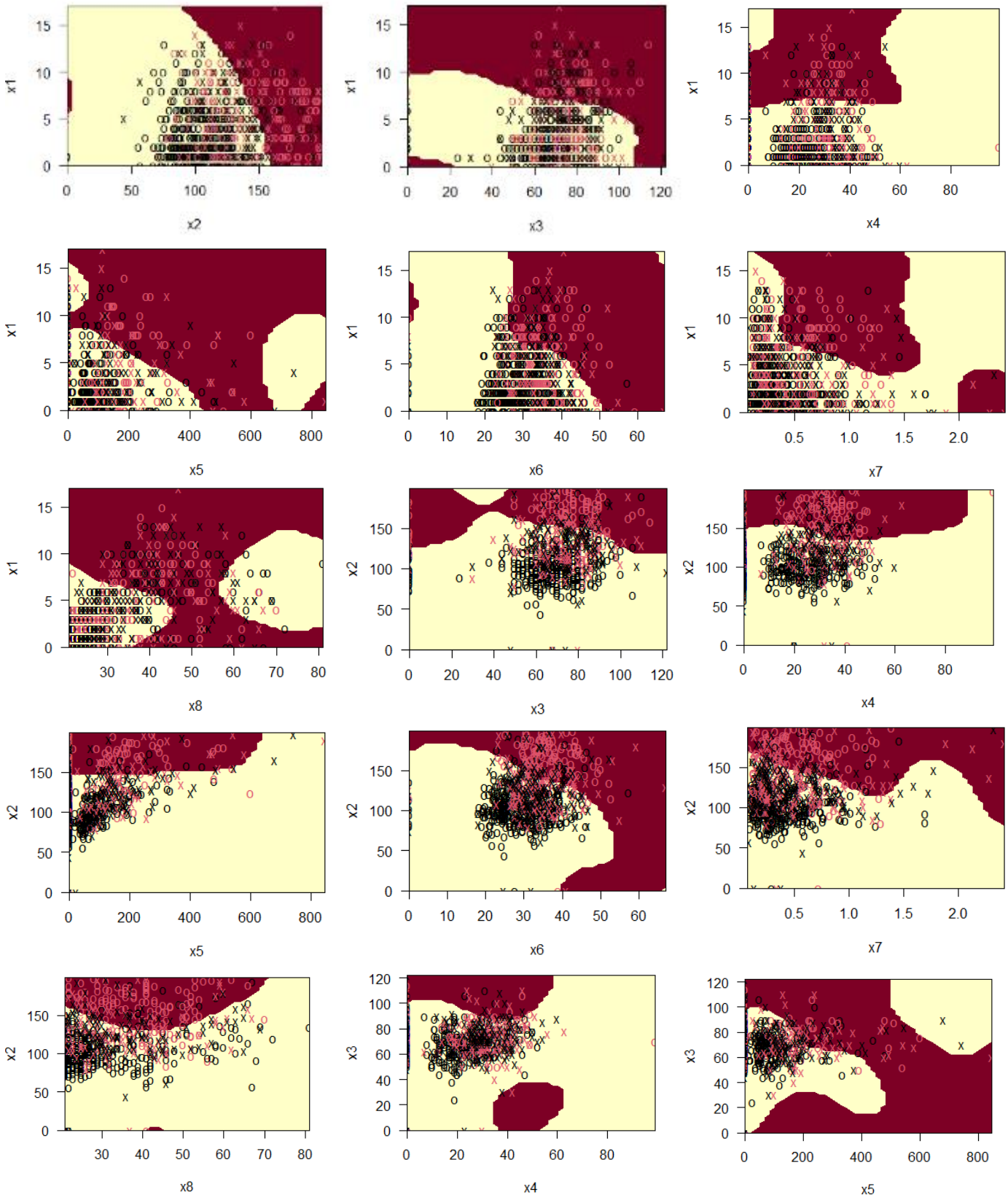
Dalam SVM terdapat beberapa fungsi kernel yang digunakan apabila data yang digunakan tidak dapat dipisahkan secara linear. Sehingga sebelum melakukan klasifikasi dengan SVM, terlebih dahulu akan dilakukan perbandingan SVM dengan empat fungsi kernel. Berikut ini tabel perbandingan fungsi kernel.

Table 4. Tingkat akurasi dari masing-masing kernel

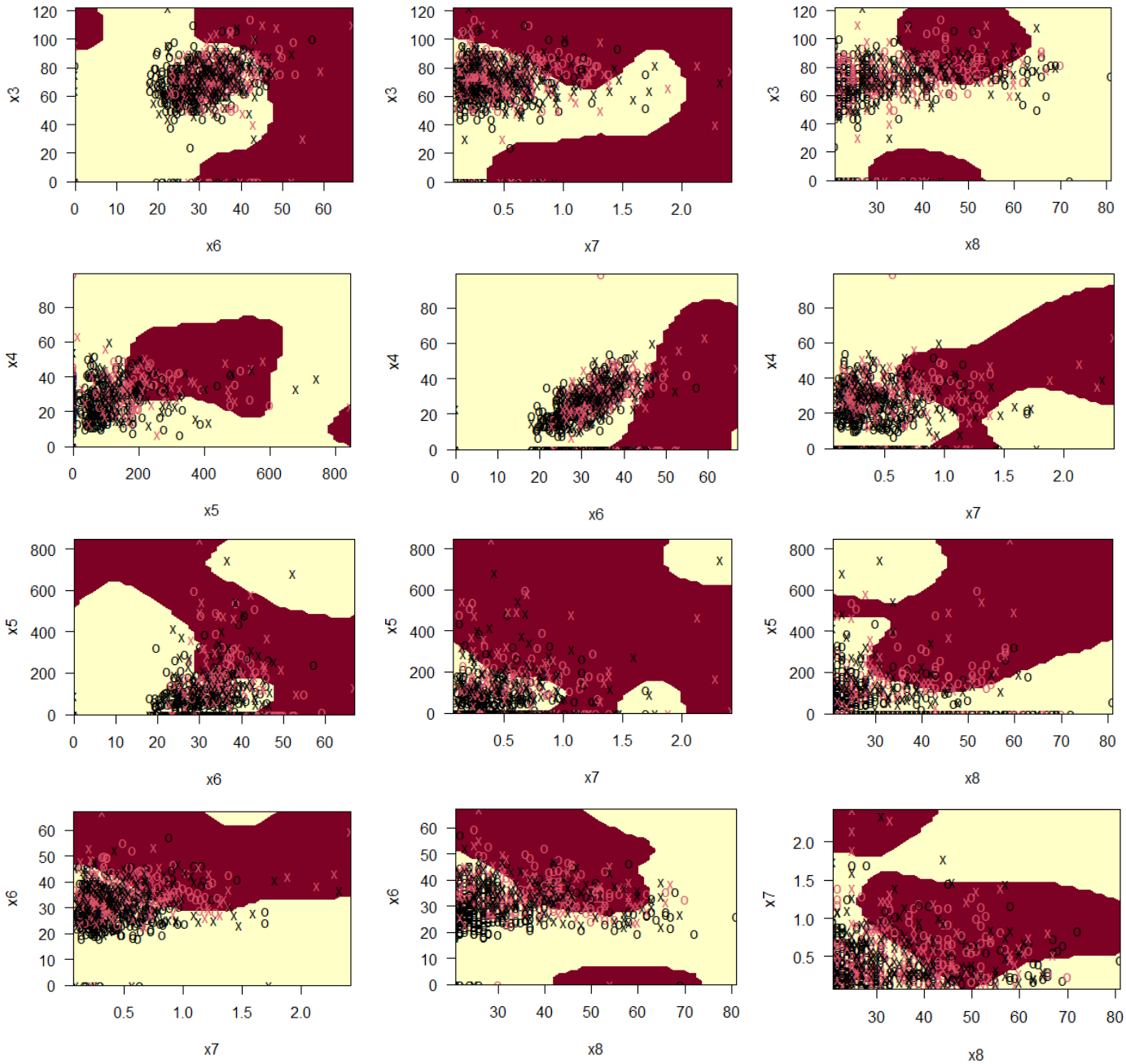
Fungsi Kernel	Tingkat Akurasi
Linear	81.13%
Polynomial	76.1%
Radial Basis	81.13%
Sigmoid	70.75%

Berdasarkan tabel diatas, dapat disimpulkan bahwa fungsi kernel yang digunakan adalah Radial Basis karena data tidak dapat dipisahkan secara linear. Gambar berikut ini menunjukkan pemisahan dengan SVM kernel radial basis.

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150



P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150



* Pregnancies (X1), Glucose (X2), Blood Pressure (X3), Skin Thickness (X4), Insulin (X5), BMI (X6), Diabetes Pedigree Function (X7), and Age (X8).

Gambar 2. Pemisahan masing-masing variabel dengan SVM

Dari gambar 2 diatas, dapat dilihat bahwa data sulit untuk dipisahkan secara linear. Hal ini karena SVM merupakan pemisahan data dalam dimensi yang cukup tinggi sehingga jika dilakukan penggambaran dalam dua dimensi maka batas pemisah datanya akan sulit dilihat. Bagian warna merah menunjukkan kelompok orang yang menderita diabetes. Sedangkan bagian warna kuning menunjukkan kelompok orang yang tidak menderita diabetes. Untuk semua pemisahan tersebut dapat dilihat bahwa data tidak dapat dipisahkan 100%. Terdapat beberapa titik kesalahan dalam pemisahan data. Dengan SVM ini kita dapat mengetahui kondisi apa yang mengakibatkan seseorang beresiko terkena diabetes. Dari gambar diatas dapat

P_ISNN: 2721-3501 e-ISSN 2721-3978
JMIK (JURNAL MAHASISWA ILMU KOMPUTER)
Vol. 4 No. 2, Oktober 2023, Hal. 137-150

dilihat bahwa jika seseorang mengalami kondisi dari beberapa variabel yang digunakan maka dapat ditentukan apakah seseorang tersebut termasuk kelompok yang beresiko terkena diabetes atau termasuk kelompok yang tidak beresiko terkena diabetes. Perhatikan tabel berikut untuk melihat berapa banyak kesalahan klasifikasi dalam data diabetes dengan metode SVM:

Table 5. Cross Table SVM

Data Testing	Klasifikasi SVM		Row Total
	No	Yes	
No	198	24	222
	0.892	0.108	0.698
	0.846	0.286	
	0.623	0.075	
Yes	36	60	96
	0.375	0.625	0.302
	0.154	0.714	
	0.113	0.189	
Column Total	234	84	318
	0.736	0.264	

Dari tabel diatas kita dapat melihat banyaknya kesalahan klasifikasinya. Pertama terlihat bahwa 198 orang yang tidak menderita diabetes tepat diklasifikan kedalam kelompok yang tidak menderita diabetes. Namun terdapat 24 orang yang tidak menderita diabetes tetapi diklasifikasikan sebagai kelompok yang menderita diabetes. Jumlah kesalahan klasifikasi tersebut tidak besar. Kemudian pada klasifikasi kelompok yang menderita diabetes, terlihat bahwa 60 orang yang menderita diabetes tepat diklasifikasikan kedalam kelompok yang menderita diabetes. Namun terdapat kesalahan klasifikasi yaitu sebanyak 36 orang yang diklasifikasikan kedalam kelompok yang tidak menderita diabetes padahal mereka menderita diabetes. Persentase kesalahan klasifikasi dalam SVM ini sama dengan kNN namun terdapat perbedaan pada posisi kesalahan klasifikasinya. Sehingga dengan demikian SVM dan kNN memiliki akurasi yang sama namun lebih baik dibanding dengan decision tree. Berikut ini adalah perbandingan ketiga metode tersebut.

Perbandingan ketiga metode mechine learning

Ketiga metode mechine learning pada dasarnya memiliki kelebihan dan kekurangan masing-masing. Namun pada artikel ini akan dilihat perbandingan ketiga metode dalam klasfikasi data diabetes untuk kedelapan variabel bebas yang digunakan. Berikut ini adalah tabel perbandingannya.

Table 6. Perbandingan Metode Machine Learning

Method	Accuracy	Sensitivity	Specificity	Balanced Accuracy
kNN	81.13%	84.91%	70.93%	77.92%
Decision Tree	72.96%	77.64%	56.94%	67.29%
SVM	81.13%	84.62%	71.43%	78.02%

Dari tabel diatas, perhatikan bahwa metode kNN dan SVM sama baiknya dalam proses pengklasifikasian dengan akurasi yang sama besar yaitu 81.13%. Tetapi dengan melihat nilai sensitivity KNN lebih baik dibanding SVM. Namun sebaliknnya dengan melihat specificity dapat disimpulkan bahwa algoritma SVM lebih baik dibanding dengan KNN karena memiliki nilai specificity yang lebih besar. Selanjutnya dengan melihat balanced accuracy maka dapat disimpulkan SVM lebih baik dibanding dengan KNN. Secara keseluruhan KNN dan SVM adalah dua metode terbaik dibanding dengan Decision Tree.

KESIMPULAN

Kesimpulan yang didapat adalah dengan diperoleh hasil akurasi yang lebih besar dari 70% untuk semua metode menunjukkan bahwa metode yang digunakan dalam artikel ini cukup baik dalam pengklasifikasian data diabetes. Tingkat akurasi yang besar juga menunjukkan bahwa variabel-variabel

yang digunakan mempunyai pengaruh terhadap kasus diabetes. Dalam artian bahwa jika diketahui kondisi seseorang berdasarkan delapan variabel yang ada maka dapat diketahui apakah orang tersebut beresiko menderia penyakit diabetes. Dalam perbandingan metode yang digunakan, diperoleh bahwa metode algorima kNN dengan k=15 lebih baik dalam proses pengklasifikasian data diabetes. Dimana SVM juga memiliki akurasi yang sama besarnya dengan KNN. Secara umum berdasarkan sensitivity, specificity, dan balanced accuracy kedua metode ini (KNN dan SVM) merupakan metode terbaik dibanding dengan Decision Tree.

REFERENSI

[1] Adem, K. (2020). Diagnosis of breast cancer with Stacked autoencoder and Subspace kNN. *Physica A: Statistical Mechanics and Its Applications*, *551*, 124591. https://doi.org/10.1016/j.physa.2020.124591

[2] Alcaraz, J., Labbé, M., & Landete, M. (2022). Support Vector Machine with feature selection: A multiobjective approach. *Expert Systems with Applications*, *204*(April), 117485. https://doi.org/10.1016/j.eswa.2022.117485

[3] Almomany, A., Ayyad, W. R., & Jarrah, A. (2022). Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), 3815–3827. https://doi.org/10.1016/j.jksuci.2022.04.006

[4] Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*(May), 100071. https://doi.org/10.1016/j.dajour.2022.100071

[5] Battineni, G., Chintalapudi, N., & Amenta, F. (2019). Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked*, *16*(June), 100200. https://doi.org/10.1016/j.imu.2019.100200

[6] Campbell, M. D., Sathish, T., Zimmet, P. Z., Thankappan, K. R., Oldenburg, B., Owens, D. R., Shaw, J. E., & Tapp, R. J. (2020). Benefit of lifestyle-based T2DM prevention is influenced by prediabetes phenotype. *Nat Rev Endocrinol*, *16*(7), 395–400. https://doi.org/doi: 10.1038/s41574-019-0316-1.

[7] Chanal, D., Steiner, N. Y., Petrone, R., Chamagne, D., & Péra, M.-C. (2022). Online Diagnosis of PEM Fuel Cell by Fuzzy C-Means Clustering. *Encyclopedia of Energy Storage*, *2*, 359–393. https://doi.org/https://doi.org/10.1016/B978-0-12-819723-3.00099-8

[8] Chen, C., Li, S. xian, Wang, S. mei, & Liang, S. wang. (2011). A support vector machine based pharmacodynamic prediction model for searching active fraction and ingredients of herbal medicine: Naodesheng prescription as an example. *Journal of Pharmaceutical and Biomedical Analysis*, *56*(2), 443–447. https://doi.org/10.1016/j.jpba.2011.05.010

[9] Chen, Y., Bai, Y., Xu, N., Zhou, M., Li, D., Wang, C., & Hu, Y. (2017). Classification of Chinese Vinegars Using Optimized Artificial Neural Networks by Genetic Algorithm and Other Discriminant Techniques. *Food Analytical Methods*, *10*, 2646–2656. https://doi.org/https://doi.org/10.1007/s12161-017-0829-y

[10] Deepa, G., Mary, G. L. R., Karthikeyan, A., Rajalakshmi, P., Hemavathi, K., & Dharanisri, M. (2022). Detection of brain tumor using modified particle swarm optimization (MPSO)

segmentation via haralick features extraction and subsequent classification by KNN algorithm. *Materials Today: Proceedings*, *56*, 1820–1826. https://doi.org/10.1016/j.matpr.2021.10.475

[11] Edla, D. R., & Cheruku, R. (2017). Diabetes-Finder: A Bat Optimized Classification System for Type-2 Diabetes. *Procedia Computer Science*, *115*, 235–242. https://doi.org/10.1016/j.procs.2017.09.130

[12] Elhazmi, A., Al-Omari, A., Sallam, H., Mufti, H. N., Rabie, A. A., Alshahrani, M., Mady, A., Alghamdi, A., Altalaq, A., Azzam, M. H., Sindi, A., Kharaba, A., Al-Aseri, Z. A., Almekhlafi, G. A., Tashkandi, W., Alajmi, S. A., Faqihi, F., Alharthy, A., Al-Tawfiq, J. A., … Arabi, Y. M. (2022). Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU. *Journal of Infection and Public Health*, *15*(7), 826–834. https://doi.org/10.1016/j.jiph.2022.06.008

[13] Freitas, M. M. dos S., Barbosa, J. R., dos Santos Martins, E. M., da Silva Martins, L. H., de Souza Farias, F., de Fátima Henriques Lourenço, L., & da Silva e Silva, N. (2022). KNN algorithm and multivariate analysis to select and classify starch films. *Food Packaging and Shelf Life*, *34*(November). https://doi.org/10.1016/j.fpsl.2022.100976

[14] Gómez-Meire, S., Campos, C., Falqué, E., Díaz, F., & Fdez-Riverola, F. (2014). Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Research International*, *60*, 230–240. https://doi.org/10.1016/j.foodres.2013.09.032

[15] Hu, A., Liao, H., Guan, W., Dong, J., & Qian, X. (2023). Journal of Radiation Research and Applied Sciences Support vector machine model based on OTSU segmentation algorithm in diagnosing bronchiectasis with chronic airway infections. *Journal of Radiation Research and Applied Sciences*, *16*(1), 100500. https://doi.org/10.1016/j.jrras.2022.100500

[16] Huang, H., Yang, Z. H., Gu, Z. W., Luo, M., & Xu, L. (2022). Decision Tree Model for Predicting the Overall Survival of Patients with Diffused Large B-Cell Lymphoma in the Central Nervous System. *World Neurosurgery*, *166*, e189–e198. https://doi.org/10.1016/j.wneu.2022.06.139

[17] Hussein, W. N., Mohammed, Z. M., & Mohammed, A. N. (2022). Identifying risk factors associated with type 2 diabetes based on data analysis. *Measurement: Sensors*, *24*(September), 100543. https://doi.org/10.1016/j.measen.2022.100543

[18] Kappelhof, N., Ramos, L. A., Kappelhof, M., van Os, H. J. A., Chalos, V., van Kranendonk, K. R., Kruyt, N. D., Roos, Y. B. W. E. M., van Zwam, W. H., van der Schaaf, I. C., van Walderveen, M. A. A., Wermer, M. J. H., van Oostenbrugge, R. J., Lingsma, H., Dippel, D., Majoie, C. B. L. M., & Marquering, H. A. (2021). Evolutionary algorithms and decision trees for predicting poor outcome after endovascular treatment for acute ischemic stroke. *Computers in Biology and Medicine*, *133*(April), 104414. https://doi.org/10.1016/j.compbiomed.2021.104414

[19] Lantz, B. (2013). *Machine Learning with R*. Packt Publishing Ltd.

[20] Lowie, T., Callens, J., Maris, J., Ribbens, S., & Pardon, B. (2021). Decision tree analysis for pathogen identification based on circumstantial factors in outbreaks of bovine respiratory disease in calves. *Preventive Veterinary Medicine*, *196*(April), 105469. https://doi.org/10.1016/j.prevetmed.2021.105469

[21] Sekaran, U., & Bougie, R. (2016). *Research Methods for Business: A Skill Building Approach*. John wiley & Sons.

[22] Shirali, G. A., Noroozi, M. V., & Maleh, A. S. (2018). Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran. *Journal of Public Health Research*, *7*, 74–80.

[23] Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

[24] Tanase, D. M., Gosav, E. M., Costea, C. F., Ciocoiu, M., Lacatusu, C. M., Maranduca, M. A., Ouatu, A., & Floria, M. (2020). The Intricate Relationship between Type 2 Diabetes Mellitus (T2DM), Insulin Resistance (IR), and Nonalcoholic Fatty Liver Disease (NAFLD). *Journal of Diabetes Research*. https://doi.org/https://doi.org/10.1155/2020/3920196

[25] Tsalera, E., Papadakis, A., & Samarakou, M. (2020). Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm. *Energy Reports*, *6*(June), 223–230. https://doi.org/10.1016/j.egyr.2020.08.045

[26] Turcato, G., Zaboli, A., Pfeifer, N., Maccagnani, A., Tenci, A., Giudiceandrea, A., Zannoni, M., Ricci, G., Bonora, A., & Brigo, F. (2021). Decision tree analysis to predict the risk of intracranial haemorrhage after mild traumatic brain injury in patients taking DOACs. *American Journal of Emergency Medicine*, *50*, 388–393. https://doi.org/10.1016/j.ajem.2021.08.048

[27] Wu, X., Wang, S., & Zhang, Y. (2017). Survey on theory and application of k-NearestNeighbors algorithm. *Comput. Eng. Appl.*, *53*(21), 1–7.

[28] Zhang, J., Wang, C., Wu, W., Jin, Q., Wu, J., Yang, L., An, Y., Yao, C., Wei, W., Song, J., Wu, W., & Guo, D. an. (2022). Authentication of herbal medicines from multiple botanical origins with cross-validation mebabolomics, absolute quantification and support vector machine model, a case study of Rhizoma Alismatis. *Arabian Journal of Chemistry*, *15*(10), 104118. https://doi.org/10.1016/j.arabjc.2022.104118

[29] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, *8*(3). https://doi.org/10.1145/2990508

[30] Zhao, D., Hu, X., Xiong, S., Tian, J., Xiang, J., Zhou, J., & Li, H. (2021). k-means clustering and kNN classification based on negative databases. *Applied Soft Computing*, *110*, 107732. https://doi.org/10.1016/j.asoc.2021.107732

[31] Zhu, Y., Wu, X., Liu, H., Niu, Z., Zhao, J., Wang, F., Mao, R., Guo, X., Zhang, C., Wang, Z., Chen, J., & Fang, Y. (2022). Employing biochemical biomarkers for building decision tree models to predict bipolar disorder from major depressive disorder. *Journal of Affective Disorders*, *308*(March), 190–198. https://doi.org/10.1016/j.jad.2022.03.080