Analisis Algoritma *Support Vector Machine* Dalam Klasifikasi Penyakit *Stroke Support Vector Machine Algorithm Analysis In Stroke Disease Classification*

1st: Kenny Riva Sulaeman
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
1aloysiusagias@student.telkomuniversity.ac.id

2nd Casi Setianingsih
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
etiacasie@telkomuniversity.ac.id

3rd Randy Erfa Saputra
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
resaputra@telkomuniversity.ac.id.ac.id

Abstrak

Stroke merupakan penyakit yang menyerang fungsional otak yang dalam istilah medis biasa disebut Transient Ischaemic Attack merupakan penyakit neurologik yang terjadi karena gangguan suplai darah menuju bagian otak yang terjadi secaramendadak. Penanganan stroke harus dilaksanakan secara cepat dan tepat guna menghindari kecacatan atau komplikasi lanjut. Di era teknologi yang sangat berkembang ini penulis membuat program machine learning guna mengidentifikasi seseorang untuk terkena stroke, agar masyarakat lebih sadar agar akan bahaya penyakit tersebut. Metode yang digunakan penulis adalah Support Vector Machine (SVM) untuk mengklasifikasi seseorang untuk terdampak penyakit stroke, metode SVM ini sangat cocok untuk digunakan karena SVM memiliki akurasi yang cukup bagus untuk sebuah klasifikasi. Dengan adanya program ini seseorang akan mengetahui seberapa persen kemungkinan untuk seseorang tersebut terkena stroke dan tidak terkena stroke, yang akan diketahui melalui persentase yang akan keluar setelah program dijalankan. Tujuan daripenulis membuat program ini adalah untuk menguji algoritma SVM dalam klasifikasi data penyakit stroke, program ini menggunakan klasifikasi SVM yang mendapatkan hasil akurasi tertinggi dari data unbalance pada kernel linear yaitu 76% dan polynomial sebesar 80%. Untuk data yang balanced penulis mendapatkan hasil akurasi pada kernel linear 77%, dan di polynomial 76%.

Kata Kunci: Stroke, Machine Learning, SupportVector Machine.

Abstract

Stroke is a functional brain disease which in medical terms is commonly called Transient Ischemic Attack, is a neurological disease that occurs due to disruption of blood supply to parts of the brain that occurs forcibly. Stroke management must be carried out quickly and appropriately to avoid or further complications. In this era of highly developed technology, the author makes a machine learning program to identify someone to have a stroke, so that people are more aware of the dangers of the disease. The method used by the author is the Support Vector Machine (SVM) to classify someone affected by stroke, this SVM method is very suitable for use because SVM has a pretty good accuracy for a classification. With this program, a person will know what percentage of the chance that someone will have a stroke and not have a stroke, which will be known through the percentage that will come out after the program is run. The purpose of the author of making this program is to test the SVM algorithm in the classification of stroke data, this program uses the SVM classification which gets the highest accuracy results from unbalance data in the linear kernel, which is 76% and polynomial by 80%. For balanced data, the authors get accuracy results in the linear kernel 77%, and in the polynomial 76%.

Keywords: Stroke, Machine Learning, SupportVector Machine.

I. PENDAHULUAN

Stroke adalah masalah kesehatan global yang umum dan serius. Di sebagian besar negara, stroke adalah yang kedua atau ketiga penyebab kematian paling umum dan salah satu penyebab utama cacat dewasa yang didapat [1]. Diperkirakan ada 4,5 juta kematian per tahun akibat stroke di dunia dan lebih dari 9 juta penderita stroke. Risiko kekambuhan selama 5 tahun adalah 15-40%. Diperkirakan pada tahun 2023 akan ada absolut peningkatan jumlah pasien yang mengalami pertama kali stroke meningkat sekitar 30% dibandingkan dengan 1983. Stoke adalah yang terdepan penyebab kecacatan pada orang dewasa [2]. Penulis ingin memanfaatkan perkembangan di bidang teknologi, yaitu metode kecerdasan buatan (AI). Tidak banyak orang yang mengetahuinya kecerdasan buatan terdiri dari beberapa cabang, salah satunya yang merupakan pembelajaran mesin atau *machine learning*. *Machine Learning* (ML) ini merupakan salah satu cabang dari AI yang sangat menarik, mengapa? Karena pembelajaran mesin adalah mesin yang bisa belajar seperti manusia. Dan di sini penulis ingin mengimplementasikan program yang dapat mengklasifikasikan penyakit stroke dengan mudah dengan machine learning dengan menggunakan metode SVM. Itu Metode Support Vector Machine (SVM) adalah salah satu mesin metode pembelajaran dengan model pembelajaran terawasi yang dapat digunakan untuk klasifikasi atau prediksi dan analisis regresi [3].

II. KAJIAN TEORI

A. Stroke

Stroke adalah serangan otak yang terjadi secara tiba-tiba, secara progresif, cepat dalam bentuk fokal atau global Defisit neurologis yang ditandai dengan gangguan aliran darah akibat penyumbatan atau pecahnya pembuluh darah di otak yang menyebabkan sel-sel otak kekurangan darah bersama dengan zat-zat yang dibawa oleh darah seperti oksigen dan makanan, dan ditandai dengan gangguan fungsi bagian tubuh fitur tubuh tertentu seperti asimetri wajah, artikulasi bicara menjadi cadel, atau lengan dan kaki menjadi lemah, yang dapat mengakibatkan kematian sel-sel tersebut dalam waktu singkat [4].

B. Machine Learning

Pembelajaran mesin atau biasa disebut *Machine Learning* (ML) adalah kategori kecerdasan buatan yang memungkinkan komputer untuk berpikir dan belajar dengan cepat secara otomatis atau pada sendiri melalui pengalaman dapat komputer. Itu juga salah satu teknologi tercepat yang berkembang untuk saat ini, dan juga mulai didukung oleh perkembangan algoritma pembelajaran dan teori [5]. Dengan ini manusia dapat membuat komputer membuat keputusan dengan memodifikasi Tindakan mereka untuk meningkatkan tindakan untuk mencapai lebih banyak akurasi, di mana akurasi diukur dalam hal berapa kali tindakan itu dipilih dan menghasilkan tindakan yang benar [6].

C. Klasifikasi

Klasifikasi adalah pengelompokan yang dilakukan sistematis pada dokumen atau buku berdasarkan hal yang sama karakteristik dan dikelompokkan ke dalam kelas atau tipe tertentu [7]. Klasifikasi juga digunakan dalam banyak hal, seperti perpustakaan. Di perpustakaan klasifikasi digunakan untuk memudahkan pembaca untuk menemukan buku yang ingin mereka temukan, karena setiap dokumen dan buku yang tersedia telah diklasifikasikan menurut alfabet dan jenis.

D. Support Vector Machine

SVM merupakan metode *machine learning* yang pertama kali diperkenalkan oleh Boser, Guyon & Vapnik pada tahun 1992 saat dipresentasikan pada *Annual Workshop on Computational Learning Theory*. SVM dapat digunakan untuk klasifikasi atau prediksi. Konsep klasifikasi dengan SVM adalah mencari apa yang disebut sebagai pemisah terbaik (*hyperplane*) antara dua kelas data. Sebuah *hyperplane* dikatakan baik jika memiliki margin terbesar. Margin adalah dua kali jarak antara *hyperplane* dengan support vector. Titik terdekat dengan hyperplane disebut support vector. SVM dapat digunakan untuk memecahkan masalah data berdimensi tinggi dan sampel pelatihan yang kecil. SVM merupakan metode yang bekerja berdasarkan prinsip *Structural Risk Minimization* (SRM). SRM digunakan untuk memaksimalkan margin dan meminimalkan batas atas risiko yang diharapkan dari risiko [8].

E. Kernel *Linear*

Kernel linier adalah kernel yang biasa digunakan untuk mengklasifikasikan data dipisahkan oleh garis yang disebut *hyperplane*. Kernel linier adalah juga merupakan kernel yang paling sederhana, dan banyak digunakan terutama untuk data representasi, kernel ini juga dapat digunakan untuk teks *mining*. Metode kernel ini dikenal sebagai klasifikasi terbaik teknik dengan cara ini memisahkan data atau memisahkan fungsi data biasa disebut *Hyperplane*. Namun, biaya pelatihan dan prediksi tinggi untuk data besar. Di sisi di sisi lain, klasifikasi linier dapat dengan mudah ditingkatkan, tetapi akan kalah jika dari segi prediktabilitas [9].

F. Kernel *Polynomial*

Kernel polinomial adalah kernel yang umum digunakan dengan Dukungan Vector Machine (SVM) dan model kernel lainnya, kernel polinomial cukup menarik, karena kernel apapun dapat ditulis sebagai nomor kernel polinomial melalui seri ekspansi Taylor. Jika kita dapat memperkirakan kernel polynomial efisien, maka kita akan dapat memperkirakan banyak secara efisien jenis kernel. Perkiraan Avron, Nguyen dan Woodruff polinomial kernel dengan matriks sketsa dalam waktu secara eksponensial ke derajat p dari kernel polinomial. Terbaru pekerjaan dari Woodruff telah meningkatkan ketergantungan pada p to polinomial [10].

G. K-Fold Cross Validation

Salah satu teknik resampling yang paling populer adalah *K-Fold Cross Validation* (KCV) adalah cara yang efisien dan andal teknik. Teknik K-Fold Cross Validation ini adalah pemisah dataset yang akan digunakan menjadi k bagian, yang akan dibagi menjadi data uji dan data latih [11].

III. METODE

A. Perancangan Sistem

Perancangan sistem ini menggunakan berbagai macam proses, seperti proses pengambilan data, *preprocessing* data, *cleaning data*, Normalisasi data, mengecek keterhubungan data (*Data Correlation*), dengan menggunakan metode *support vector machine*. Program ini dapat membantu seseorang untuk mendapatkan tingkat akurasi orang tersebut untuk terkena penyakit stroke.

B. Preprocessing

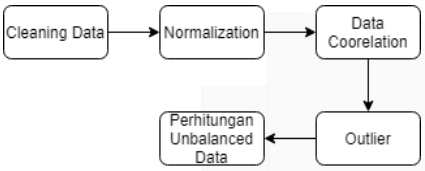
Preprocessing adalah teknik awal data mining untukmengubah data mentah yang dikumpulkan dari berbagai sumber menjadi lebih bersih dan informasi yang lebih mudah diakses yang digunakan untuk pemrosesan lebih lanjut, yang seringkali dapat memberikan dampak yang signifikan terhadap kinerja generalisasi pembelajaran mesin yang diawasi algoritma. Dari total 5000 baris dan 12 kolom data di dataset, penulis hanya mengambil 1000 baris dan 12 kolom data untuk pengujian. Di sini penulis melakukan dua tes pada data, yang pertama adalah data yang tidak seimbang dengan total 1000 data dengan yang tidak terpengaruh berjumlah 751 data sedangkan yang terpengaruh adalah 249 data, dan data seimbang adalah 500 data tanpa data yang terpengaruh adalah 251 dan yang terkena stroke adalah 249. seperti pada tabel di bawah:

Tabel 1. *Unabalanced Data*

index	ID	Gender	Age	Hypertension	Heart Disease	Ever Married	Work Type	Residence Type	Avg Glucose Level	BMI	Smoking Status	Stroke
0	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
…	..	…	…	…	…	…	…	…	…	…	…	…
995	60211	Male	1.4	0	0	No	Children	Urban	90.51	18.9	unknown	0
996	53279	Male	0.24	0	0	No	Children	Rural	118.87	16.3	unknown	0
997	61715	Male	55	0	0	Yes	Private	Rural	56.42	31.8	never smoked	0
998	37830	Female	29	0	0	No	Private	Urban	73.67	21	unknown	0
999	2454	Male	4	0	0	No	Children	Rural	89.11	20.1	unknown	0
index	ID	Gender	Age	Hypertension	Heart Disease	Ever Married	Work Type	Residence Type	Avg Glucose Level	BMI	Smoking Status	Stroke
0	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1

3	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
…	..	..	…	…	…	…	…	…	…	…	…	..
995	60211	Male	1.4	0	0	No	Children	Urban	90.51	18.9	unknown	0
996	53279	Male	0.24	0	0	No	Children	Rural	118.87	16.3	unknown	0
997	61715	Male	55	0	0	Yes	Private	Rural	56.42	31.8	never smoked	0
998	37830	Female	29	0	0	No	Private	Urban	73.67	21	unknown	0
999	2454	Male	4	0	0	No	Children	Rural	89.11	20.1	unknown	0

Gambar 1 *Preprocessing Data*



Data *preprocessing* berguna agar data bisa diproses ke langkah berikutnya yaitu dengan metode SVM.

F. SVM *Clasification*

Metode klasifikasi yang digunakan dalam penelitian ini adalah metode *Support Vector Machine* (SVM), Metode SVM adalah sangat dianjurkan karena pada penelitian sebelumnya telah terbukti mendapatkan akurasi yang tinggi. Metode SVM sangat cocok untuk penelitian ini karena dataset yang digunakan berisi bilangan biner. Disini penulis menggunakan 2 jenis kernel, yaitu kernel *linear* dan *polynomial*.

G. Kernel *Linear*

Di kernel linier, penulis melakukan berbagai eksperimen. Di kernel linier penulis mengubah C, Gamma dan *Cross Validation* berbeda. Hal ini bertujuan untuk mendapatkan beberapa hasil berbeda, C = [0.1, 1, 10, 100], Gamma = [0,01, 0,1, 1, 10, 100] dan CV = [3, 5, 7]. Pada kernel linier dilakukan mulai dari C = 0,1, Gamma = 0,01 dan CV = 3 sampai C = 100, Gamma = 100, dan CV = 7, dengan total 60 percobaan pada kernel linier.

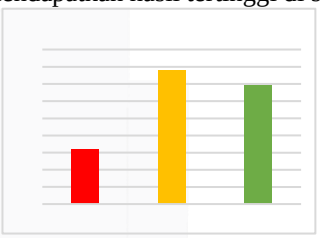
H. Kernel *Polynomial*

Dalam kernel polinomial penulis melakukan percobaan pada dataset, dengan mengubah derajat di kernel untuk mendapatkan hasil yang berbeda juga. Dan dalam menguji ini program yang penulis gunakan *Degree* = [2, 3, 4, 5, 6, 7] dan CV = [3, 5, 7]. Anda dipersilakan seperti kernel linier penulisnya bereksperimen dengan mengubah angka pada *Degree* dan CV mulai dari *Degree* = 2 dan CV = 3 hingga *Degree* = 7 dan CV = 7, dengan total 18 percobaan pada kernel polinomial dengan angka yang berbeda.

IV. HASIL DAN PEMBAHASAN

A. Kernel *Linear K-Fold Cross Validation Unbalanced Data*

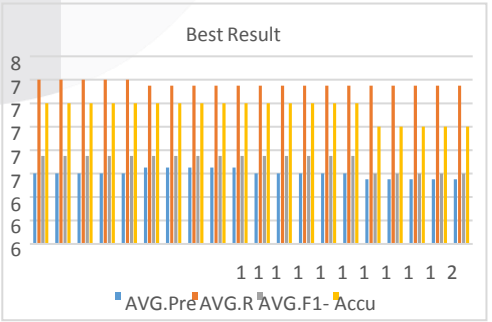
Penulis melakukan sebanyak K kali dengan K berbeda-beda, pengujian dilakukan secara bertahap dimulai dari K = 3, K = 5, K = 7, dan juga dengan C dan Gamma berbeda-beda, kemudian akan dicari nilai rata-rata akurasi terbesar dari masing-masing K. Dan mendapatkan hasil tertinggi di saat K = 5.



Gambar 2 *K-Fold Cross Validation* Kernel *Linear*

B. Pengujian Kernel *Linear Unbalanced Data*

Di kernel linier penulis melakukan berbagai eksperimen dengan mengubah angka dalam C dan Gamma dalam program. C = (0,1, 1, 10, 100) Gamma = (0,01, 0,1, 1, 10, 100) dan K = (3, 5, 7), dengan total uji coba sebanyak 60 kali

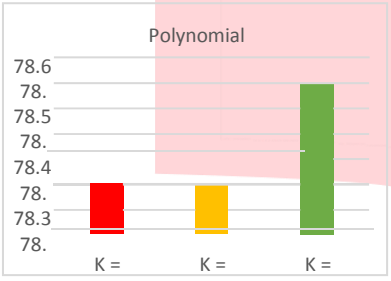


Gambar 3. Hasil Terbaik Kernel *Linear Unbalanced Data*

Dari semua pengujian di ambil data dengan hasil terbaik mulai dari rata-rata *precision*, *recall*, *F1Score*, dan *Accuracy*, lalu dijadikan diagram batang seperti gambar diatas.

C. Kernel *Polynomial K-Fold Cross Validation Unbalanced Data*

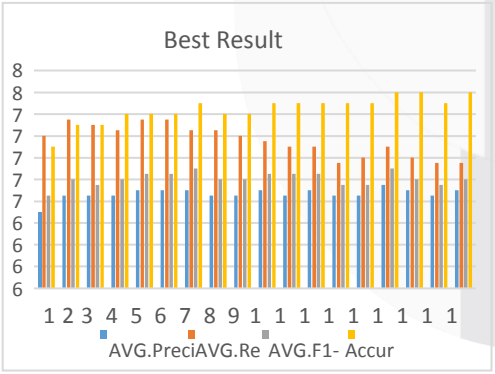
Pada kernel *polynomial* penulis juga melakukan sebanyak K kali dengan K berbeda-beda, pengujian dilakukan secara bertahap dimulai dari K = 3, K = 5, dan K = 7, kemudian akan dicari nilai rata-rata akurasi terbesar dari masing-masing K. Untuk kernel *polynomial* penguji mengubah degreenya dengan Degree = [2,3,4,5,6,7]. Dan akurasi tertinggi di dapat saat K = 7.



Gambar 4 Hasil Terbaik *K-Fold Cross Validation* Kernel *Polynomial*

D. Pengujian Kernel *Polynomial Unbalanced Data*

Pada data yang tidak seimbang, pengujian dilakukan dengan kernel polinomial, dan penulis melakukan berbagai eksperimen dengan mengubah degree dan cross validation pada kernel. Degree = (2, 3, 4, 5, 6, 7) dan CV = (3, 5, 7) dengan total 18 percobaan.



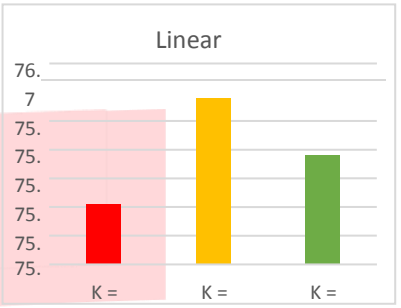
Gambar 5 Hasil Terbaik Kernel *Polynomial Unbalanced Data*

Dari semua pengujian di ambil data dengan hasil terbaik mulai dari rata-rata precision, recall, F1-Score, dan Accuracy, lalu dijadikan diagram batang seperti gambar diatas

E. Kernel *Linear K-Fold Cross Validation Balanced Data*

Pada data berimbang ini, penulis melakukan hal yang sama seperti

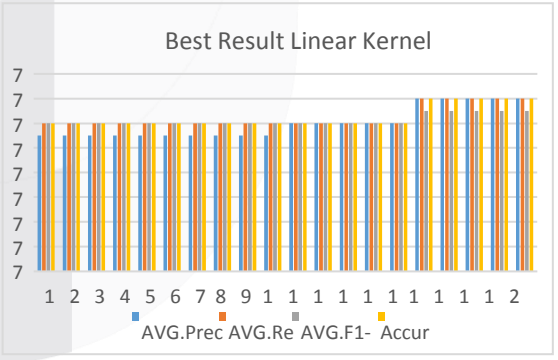
Sebelumnya, percobaan dilakukan K kali, dengan K yang berbeda, pengujian dilakukan secara bertahap mulai dari K = 3, K = 5, dan K = 7, maka akan dicari nilai rata-rata dari akurasi terbesar dari setiap K dan dapatkan rata-ratanya akurasi seperti yang ditunjukkan di bawah ini.



Gambar 6 Hasil Terbaik *K-Fold Cross Validation* Kernel *Linear Balanced Data*

F. Pengujian Kernel *Linear Balanced Data*

Pada data berimbang ini, penulis melakukan hal yang sama seperti Sebelumnya, percobaan dilakukan K kali, dengan K yang berbeda, pengujian dilakukan secara bertahap mulai dari K = 3, K = 5, dan K = 7, maka akan dicari nilai rata-rata dari akurasi terbesar dari setiap K dan dapatkan rataratanya akurasi seperti yang ditunjukkan di bawah ini.

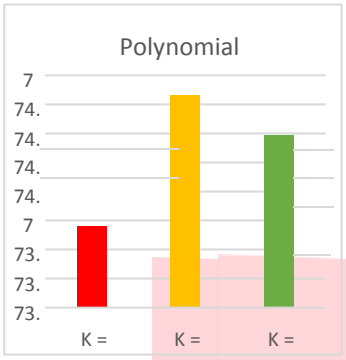


Gambar 7 Hasil Terbaik Kernel *Linear Balanced Data*

Dari semua pengujian di ambil data dengan hasil terbaik mulai dari rata-rata precision, recall, F1-Score, dan Accuracy, lalu dijadikan diagram batang seperti gambar diatas.

G. Kernel *Polynomial K-Fold Cross Validation Balanced Data*

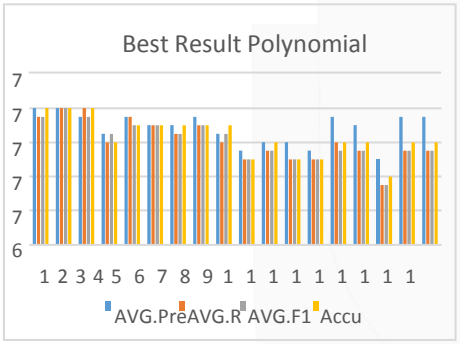
Dalam kernel polinomial, percobaan K kali, dengan K berbeda, pengujian dilakukan mulai dari K = 3, K = 5, dan K = 7, maka nilai rata-rata terbesar akan dicari akurasi dari setiap akurasi K. Akurasi tertinggi terdapat pada saat K = 5.



Gambar 4.7 Hasil Terbaik *K-Fold Cross Validation* Kernel *Polynomial*

H. Pengujian Kernel *Polynomial Balanced Data*

Pada kernel polinomial penulis melakukan percobaan dengan mengubah degree dan *cross validation* pada kernel, Degree = (2, 3, 4, 5, 6, 7) dan CV = (3, 5, 7) dengan total 18 waktu percobaan.



Gambar 4.8 Hasil Terbaik Kernel *Polynomial Balanced Data*

V. KESIMPULAN

Berdasarkan penelitian dan pengujian yang telah penulis lakukan pada Tugas Akhir ini, maka dapat disimpulkan bahwa, dalam penelitian ini penulis menggunakan dataset dari Kaggle yang diterbitkan pada tahun 2021, menggunakan Machine Learning menggunakan metode Support Vector Machine untuk mengklasifikasi data. Penulis melakukan pengujian data dengan dua cara, yaitu menggunakan teknik Unbalanced Data untuk melakukan penelitian dengan data yang tidak sebanding, dan cara klasifikasi pada umumnya dengan data sebanding (Balanced). Untuk metode Support Vector Machine penulis menggunakan dua jenis kernel yaitu kernel Linear dan Polynomial, untuk data unbalanced pada kernel linear mendapatkan akurasi terbesar 76% dan untuk polynomial mendapatkan hasil akurasi 80%. Untuk data yang balanced dengan kernel linear penulis mendapatkan hasil akurasi tertinggi 77%, sedangkan pada kernel polynomial penulis mendapatkan hasil tertinggi 76%.

REFERENSI:

[1] Wayunah, W., & Saefulloh, M. (2017). Analisis faktor yang berhubungan dengan kejadian stroke di rsud indramayu. Jurnal Pendidikan Keperawatan Indonesia, 2(2), 65-76.

[2] Wolfe, C. D. (2000). The impact of stroke. British medical bulletin, 56(2), 275- 286.

[3] Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). Support vector machine. Proceeding Indones. Sci. Meeiting Cent. Japan.

[4] Adelina, V., Ratnawati, D. E., & Fauzi, M. A. (2018). Klasifikasi Tingkat Risiko Penyakit Stroke Menggunakan Metode GA-Fuzzy Tsukamoto. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

[5] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[6] Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In Journal of physics: conference series (Vol. 1142, No. 1, p. 012012). IOP Publishing.

[7] Hamakonda, T. P. (1978). Pengantar klasifikasi persepuluhan Dewey. BPK Gunung Mulia.

[8] Hasanah, S. (2018). Propensity Score Matching Menggunakan Support Vector Machine Pada Kasus Diabetes Melitus (DM) Tipe 2 (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).

[9] Huang, H. Y., & Lin, C. J. (2016, June). Linear and kernel classification: When to use which?. In Proceedings of the 2016 SIAM International Conference on Data Mining (pp. 216-224). Society for Industrial and Applied Mathematics.

[10] Song, Z., Woodruff, D., Yu, Z., & Zhang, L. (2021, July). Fast sketching of polynomial kernels of polynomial degree. In International Conference on Machine Learning (pp. 9812-9823). PMLR.

[11] Anguita, D., Ghio, A., Ridella, S., & Sterpi, D. (2009, July). K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines. In DMIN (pp. 291-297).