PENGARUH KOMPOSISI *SPLIT DATA* PADA AKURASI KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN ALGORITMA *MACHINE LEARNING*

Febby Refindha Aftha Harianto[1], Zakki Alawi[2], Ita Aristia Sa'ida[3]
Program Studi Teknik Informatika[1,3]
Program Studi Sistem Informasi[2]
Universitas Nahdlatul Ulama Sunan Giri
Jl. Ahmad Yani No. 10, Sukorejo, Kabupaten Bojonegoro
e-mail: *[1]**refibjn86@gmail.com,** [2]zakki.alawi@unugiri.ac.id,
[3]itaaristia@unugiri.ac.id

Abstract
The increasing number of people with diabetes is an international health problem. To prevent diabetic complications, early diagnosis and accurate classification are essential. This study looks at how the composition of split data affects the classification performance of diabetics with machine learning algorithms such as Random Forest, Naive Bayes, and Support Vector Machine (SVM). The research data is taken from Bojonegoro Regency Hospital, which consists of 128 samples that have 10 main features. To ensure the data is ready for use, the research method goes through a preprocessing stage. Next, the data was divided into training and testing data with a ratio of 90:10, 80:20, 70:30, 60:40, and 50:50 respectively. Using confusion matrix, the algorithm is assessed for accuracy, precision, recall, and F1 score. In this study we focus on the accuracy values obtained and the results show that the proportion of data sharing affects the performance of the algorithm. Random Forest achieved 100% accuracy in some scenarios. This algorithm also proved to be the most effective in the classification of diabetics. In conclusion, algorithm selection and data split composition are very important for model performance optimization. These results are important for the development of more accurate and efficient Machine Learning-based diagnosis systems. Further research can consider larger datasets and additional algorithms for better results.

Keyword: *Algorithm, Classification, Diabetes, Machine Learning, Split Data*

PENDAHULUAN

Jumlah orang yang menderita diabetes diperkirakan mencapai 120 juta jiwa di seluruh dunia, menjadikan diabetes menjadi salah satu masalah utama bagi dunia kesehatan global. Jika masyarakat umum tidak memahami faktor yang dapat menyebabkan diabetes, jumlah penderitanya diprediksi akan semakin bertambah (Ucha Putri et al., 2021). Salah satu tanda seseorang menderita diabetes adalah tingginya kadar gula darah dalam urine dikarenakan metabolisme tubuh terganggu yang disebabkan oleh produksi dan fungsi hormon insulin tidak berjalan dengan baik (Angriani & Baharuddin, 2020). Komplikasi fisik seperti hipertensi, kerusakan mata, kerusakan ginjal, penyakit jantung, dan stroke juga dapat terjadi karena kadar gula dalam darah yang melebihi batas normal (Yusnita et al., 2021). Deteksi dini seharusnya mampu mengurasi risiko terjadinya komplikasi bagi penderita diabetes di kemudian hari (Fathurahman et al., 2023). Keterlambatan dalam proses diagnosis terhadap pasien dapat memperburuk penanganan medis yang diterima oleh pasien tersebut (Sanjaya et al., 2023). Segera melakukan analisis pada pasien diabetes akan memungkinkan penyakit dapat dikenali lebih awal sehingga dapat mencegah seseorang terjangkit penyakit tersebut. Pada penelitian kali ini menggunakan data penderita diabetes sebanyak 128 orang yang terkena diabetes pada bulan Agustus 2023 di RSUD Kabupaten Bojonegoro, data tersebut berisi 10 fitur dengan diagnosa sebagai kelas.

Berdasarkan uraian masalah di atas data mining khususnya klasifikasi adalah metode yang dirasa dapat mengatasi masalah tersebut. Data mining adalah proses berulang yang didasarkan pada penemuan yang dilakukan secara otomatis atau manual (Terbuka et al., 2024). Sementara itu penelitian lain menyebutkan bahwa data mining merupakan sebuah proses yang dilakukan untuk mengekstrak pola atau informasi yang berguna dari kumpulan data menggunakan algoritma tertentu (Ramon et al., 2022). Salah satu isu utama dalam bidang data mining adalah klasifikasi. Dalam proses klasifikasi, sebuah model pengklasifikasi dibangun berdasarkan kumpulan data yang telah dilatih dengan kategori yang telah ditentukan sebelumnya (Munir et al., 2024). Klasifikasi banyak digunakan dalam berbagai sistem, seperti prediksi penjualan, sistem diagnosa medis, dan sebagainya. Kualitas sistem klasifikasi juga sangat dipengaruhi oleh pemilihan metode klasifikasi yang tepat (Nur Azizah et al., 2023).

Untuk membantu manusia dalam mengambil keputusan dalam berbagai hal misalnya dalam bidang medis, pendidikan, dan lain lain berbagai metode dalam *machine learning* seperti halnya K-NN, Naïve Bayes, SVM, Regresi Linier, dan lain lain dirasa dapat memberikan manfaat dalam mengolah data (Sanjaya et al., 2023). Menurut penelitian yang dilakukan oleh (Fathurahman et al., 2023), perbandingan antara dua algoritma dalam klasifikasi penderita diabetes yaitu Naive Bayes dan Random Forest menunjukan bahwa Random Forest lebih unggul dengan selisih akurasi sebesar 1.67%. Penelitian lain menyebutkan, bahwa akurasi pada algoritma SVM lebih unggul dibanding dengan algoritma Naïve Bayes dalam pengklasifikasian suatu model (Prasetyo et al., 2024).

Tujuan penelitian ini yaitu berfokus pada membandingkan dan mengukur tingkat akurasi dari ketiga metode yang digunakan dalam penelitian sebelumnya. Karena menurut penelitian yang dilakukan oleh (Ainurrohma, 2021) tingkat akurasi dapat mempengaruhi performa algoritma dalam mengolah data yang ada. Selain itu, penelitian ini juga bertujuan untuk menentukan proporsi *split data* yang paling akurat dalam penerapan ketiga metode tersebut.

METODE PENELITIAN

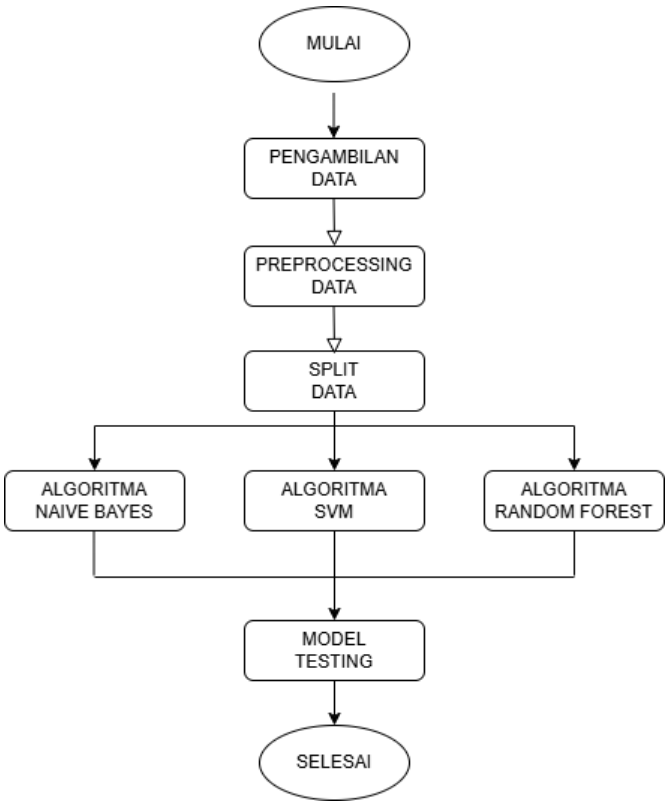
Naive Bayes adalah metode pengklasifikasian yang berbasis probabilitas sederhana, yang menghitung serangkaian probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang tersedia (Purnomo et al., 2020).

Random Forest merupakan salah satu metode klasifikasi yang paling tepat untuk melakukan prediksi, karena mampu mengelola banyak variabel tanpa mengalami overfitting. Selain itu, metode ini juga dapat mengidentifikasi hubungan antar random forest, mirip dengan karakteristik metode *ensemble* (Prasetyo et al., 2024).

Untuk menyelesaikan masalah regresi dan klasifikasi, dapat menggunakan algoritma *Support Vector Machine (SVM)* secara luas. Algoritma ini merupakan pengembangan dari *support vector classifier*, yang terkenal sebagai *classifier* dengan margin tertinggi. Dirancang untuk menangani data sederhana, algoritma ini dapat dipisahkan secara linier guna memaksimalkan margin antara dua kelas (Baiq Nurul Azmi et al., 2023).

Gambar 1 menggambarkan *flowchart* yang menunjukkan seluruh langkah dalam proses penelitian. Proses penelitian dimulai dengan pengumpulan data dari 128 pasien diabetes di RSUD Kabupaten Bojonegoro pada bulan Agustus 2023, yang mencakup 10 fitur dengan diagnosis sebagai kelas. Tahapan selanjutnya adalah *Preprocessing Data*, pada proses ini dilakukan beberapa proses pada data seperti menghapus data yang tidak akurat, menghapus data duplikasim serta merubah data ke dalam bentuk numerik agar mampu diolah secara baik oleh algoritma *Machine Learning*. Selanjutnya, dilakukan proses *splitting data*, di mana kumpulan data dibagi menjadi data latih dan data uji dengan menggunakan rasio perbandingan 90:10, 80:20, 70:30, 60:40, dan 50:50 (Prastyo et al., 2020).

P-ISSN: 2622-6901
E-ISSN: 2622-6375



Gambar 1. Tahapan Penelitian

Setelah melalui proses *split data*, langkah selanjutnya adalah menerapkan tiga algoritma Machine Learning dalam penelitian ini, yaitu Algoritma Naïve Bayes, Algoritma Support Vector Machine (SVM), dan Algoritma Random Forest

Setelah data diuji dengan ketiga metode tersebut, tahap berikutnya adalah Model Testing, di mana model data akan dievaluasi menggunakan confusion matrix. Confusion matrix merupakan alat yang digunakan untuk mengevaluasi kinerja algoritma machine learning, yang memberikan informasi tentang klasifikasi dan prediksi yang sebenarnya. Di dalamnya terdapat empat indikator yang diukur, yaitu *accuracy, precision, recall dan F1-Score* (Prastyo et al., 2020).

Tabel 1. Skenario Confusion Matrix

		Predicted Class	
		Class = True	**Class = False**
Actual Class	**Class = True**	True Positive	False Negative
	Class = False	False Positive	True Negative

Perhitungan dari 4 indikator tersebut adalah :

$$Accuracy: Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (1)$$

$$Recall: Recall = \frac{TP}{(TP+FN)} \quad (2)$$

$$Precision: Precision = \frac{TP+TN}{(TP+FP)} \quad (3)$$

$$F1 - Score: F1 - Measure = 2\ X \frac{precission\ x\ recall}{precission\ x\ recall} \quad (4)$$

Penjelasan :

- True Positive (TP) yaitu pasien yang memang memiliki diabetes (positif) dan model juga memprediksi mereka terkena diabetes (positif).
- False Positive (FP) adalah pasien yang sebenarnya tidak terkena diabetes (negatif), tetapi model salah memprediksi mereka memiliki diabetes (positif).
- False Negative (FN) yaitu pasien sebenarnya memiliki diabetes (positif), tetapi model salah memprediksi mereka tidak memiliki diabetes (negatif).
- True Negative (TN) adalah pasien yang memang tidak memiliki diabetes (negatif) dan model juga memprediksi mereka tidak memiliki diabetes (negatif).

Dari proses ini akan ditemukan rasio *split data* mana yang mampu meningkatkan akurasi dari Algoritma Naïve Bayes, SVM, dan Random Forest pada klasifikasi penderita diabetes. Setelah semua tahapan selesai dijalankan, maka akan muncul nilai akurasi dan juga rasio *split data* yang paling akurat pada setiap algoritma yang telah digunakan.

a. Pengambilan Data

Tahapan pertama kali yang dilakukan adalah mencari dan mengumpulkan dataset penderita diabetes dari RSUD Kabupaten Bojonegoro. Dataset diabetes berisi 10 variabel yang dibagi menjadi 9 fitur dan 1 kelas, dengan total 128 data penderita diabates pada bulan Agustus 2023. Detail atribut terdapat pada Tabel 1 dan 2.

Tabel 2. Fitur Dataset

No	Feature	
	Atribut	**Description**
1	jenis_kelamin	Menunjukkan jenis kelamin penderita diabetes, dengan 1=Laki-Laki, dan 2=Perempuan
2	usia	Menunjukkan usia penderita diabetes
3	IMT	Indeks Massa Tubuh (IMT) dihitung dengan cara membagi berat badan dengan kuadrat dari tinggi badan
4	tekanan_darah	Menunjukkan nilai dari tekanan darah
5	kadar_gula	Menunjukkan jumlah kadar gula dalam darah atau tubuh
6	penyakit_penyerta	Menunjukkan adanya penyakit penyerta yang diderita, dengan 0=tidak terdapat penyakit penyerta, dan 1=terdapat penyakit penyerta
7	nyeri	Menunjukkan adanya nyeri yang dirasakan, dengan 0=tidak terdapat rasa nyeri yang dirasakan, dan 1=terdapat nyeri yang dirasakan
8	riwayat_kel	Menunjukkan adanya riwayat keluarga yang pernah terkena diabetes, dengan 0=tidak terdapat riwayat dalam keluarga, dan 1=terdapat riwayat dalam kelurga
9	diet	Menunjukkan adanya usaha diet yang dilakukan, dengan 0=tidak menjalani diet, dan 1=menjalani diet

Tabel 3. Kelas Dataset

No	Class	
	Atribut	**Description**
1	diagnosa	Menunjukkan diagnosa, dengan 0=tidak terdiagnosa diabetes, dan 1=terdiagnosa terkena diabetes

P-ISSN: 2622-6901
E-ISSN: 2622-6375

b. Preprocessing Data

Dari seluruh tahapan *preprocessing* hingga data dapat diolah secara optimal oleh algoritma Machine Learning, kami akan menampilkan hasil dari proses *preprocessing* yang ditampilkan pada Gambar 3.

No	diagnosa	jenis_kelamin	usia	IMT	tekanan_darah	kadar_gula	penyakit_penyerta	nyeri	riwayat_kel	diet
1	0	2	71	20	111	78	0	0	0	1
2	0	2	63	23	96	90	0	0	1	1
3	0	2	70	25	146	95	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...
126	1	2	54	25	223	486	1	1	1	1
127	1	2	54	28	115	498	1	1	1	1
128	1	1	60	25	120	559	1	0	1	1

Gambar 2. Data setelah melewati tahap *preprocessing*

c. Split Data

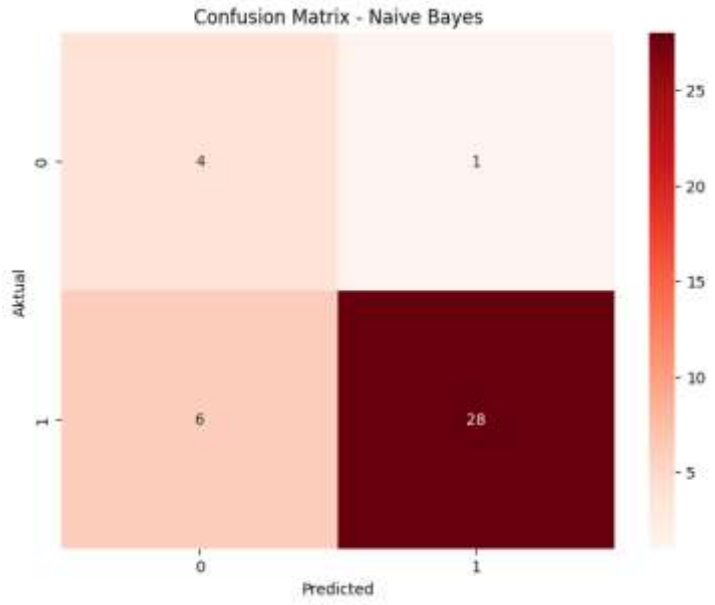
Penentuan jumlah data pelatihan dengan perbandingan rasio untuk setiap diagnosis diabetes ditentukan oleh rasio yang digunakan dalam proses *split data*, yaitu 90:10, 80:20, 70:30, 60:40, dan 50:50. Hasil akurasi dari masing-masing rasio tersebut bervariasi satu sama lain.

Tabel 4. Rasio *Splitting Data*

Data	Rasio				
	90:10	**80:20**	**70:30**	**60:40**	**50:50**
Latih	115	102	89	76	64
Uji	13	26	39	52	64

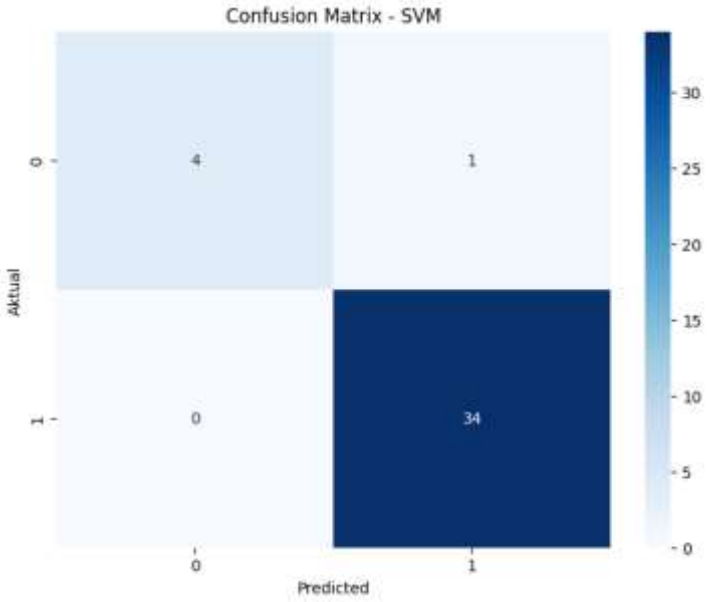
d. Model Testing

Tahap terakhir adalah model testing atau pengujian model, yang melibatkan evaluasi algoritma setelah implementasi model algoritma Naïve Bayes, SVM, dan Random Forest. Pada tahap ini, dilakukan pengukuran kinerja model algoritma Machine Learning yang telah diterapkan dengan memanfaatkan confusion matrix.

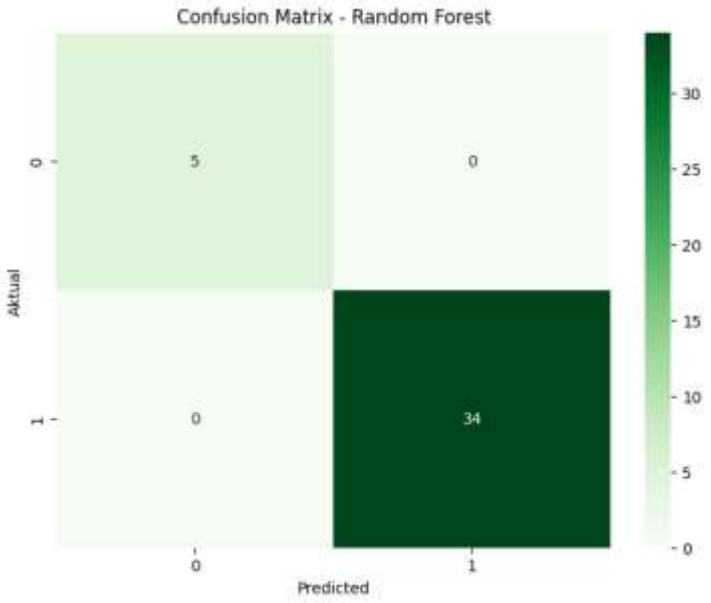


Gambar 3. Algoritma Naïve Bayes

P-ISSN: 2622-6901
E-ISSN: 2622-6375



Gambar 4. Algoritma SVM



Gambar 5. Algoritma Random Forest

HASIL DAN PEMBAHASAN

Hasil penelitian ini diperoleh dari kelima skenario yang dilakukan dalam penelitian ini dengan memanfaatkan algoritma Naïve Bayes, SVM, dan Random Forest. Tabel 5 menunjukkan hasil akurasi dari *split data* yang menggunakan Naïve Bayes, sementara Tabel 6 menampilkan hasil akurasi dari *split data* menggunakan SVM. Sedangkan pada Tabel 7 menyajikan hasil dari *split data* yang diterapkan dengan algoritma Random Forest.

Tabel 5. Algoritma Naïve Bayes

NO	Split Data		Naïve Bayes			
	Data Latih (%)	Data Uji (%)	Accuracy	Precision	Recall	F1 Score
1	90 : 10		92.30%	100%	92.30%	96%
2	80 : 20		96.15%	96.92%	96.15%	96.32%
3	70 : 30		82.05%	89.30%	82.05%	84.33%
4	60 : 40		84.61%	91.05%	84.61%	86.64%
5	50 : 50		85.93%	91.54%	85.93%	87.60%

Tabel 6. Algoritma SVM

NO	Split Data		SVM			
	Data Latih (%)	Data Uji (%)	Accuracy	Precision	Recall	F1 Score
1	90 : 10		100%	100%	100%	100%
2	80 : 20		96.15%	96.32%	96.15%	95.92%
3	70 : 30		97.43%	97.50%	97.43%	97.31%
4	60 : 40		96.15%	96.15%	96.15%	96.15%
5	50 : 50		96.87%	96.87%	96.87%	96.87%

Tabel 7. Algoritma Random Forest

NO	Split Data		Random Forest			
	Data Latih (%)	Data Uji (%)	Accuracy	Precision	Recall	F1 Score
1	90 : 10		100%	100%	100%	100%
2	80 : 20		88.46%	89.84%	88.46%	85.36%
3	70 : 30		100%	100%	100%	100%
4	60 : 40		98.07%	98.11%	98.07%	97.99%
5	50 : 50		96.87%	96.98%	96.87%	96.67%

Pada penelitian kali ini didapatkan hasil perbandingan dari ketiga algoritma yang digunakan, selain itu proporsi *split data* terbukti cukup berpengaruh untuk menghasilkan akurasi dari metode *Machine Learning* dalam mengolah data dalam hal ini data diabetes. Berdasar pada tujuan awal penelitian ini dilakukan tentunya dengan metodologi yang sudah dilakukan berjalan cukup efektif dalam menentukan algoritma dan rasio *split data* yang memiliki akurasi paling tinggi dalam klasifikasi penderita diabetes.

KESIMPULAN

Berdasarkan pengujian yang dilakukan dengan menggunakan berbagai rasio pada tahap *splitting data*, ditemukan bahwa Algoritma *Supervised Learning* dapat mengklasifikasikan penderita diabetes dengan tingkat akurasi yang cukup tinggi. Meskipun terdapat variasi dalam hasil akurasi, presisi, dan recall di antara klasifikasi yang berbeda, secara keseluruhan, Algoritma *Supervised Learning* menunjukkan hasil yang konsisten dalam mengenali dan membedakan penderita diabetes. Hasil penelitian menunjukan penggunaan Algoritma *Machine Learning* pada penelitian kali ini khususnya dalam hal ini Algoritma *Supervised Learning* dapat memberikan hasil yang cukup menjanjikan untuk mencari akurasi tertinggi dalam klasifikasi penderita diabetes.

Dari hasil skenario percobaan yang dilakukan pada penelitian kali ini didapatkan kesimpulan bahwa nilai rata-rata Algoritma Random Forest cukup unggul dibanding dengan Naïve Bayes dan SVM, dengan melihat hasil akurasi yang cukup tinggi dengan percobaan beberapa rasio pada tahap *Split Data*. Dengan ini Algoritma Random Forest terbukti mampu memberikan akurasi terbaik dalam mengidentifikasi penderita diabetes berdasarkan beberapa fitur dan bahkan akurasi mencapai 100% meskipun dirasa data yang digunakan kurang stabil.

SARAN

Berdasarkan temuan dari penelitian ini, diperlukan penelitian lanjutan untuk memperdalam pemahaman dan memperluas analisis yang telah dilakukan. Salah satu pendekatan untuk mencapainya adalah dengan menguji beberapa algoritma yang telah digunakan pada dataset yang lebih besar dan memiliki fitur yang lebih bervariasi, serta mengeksplorasi algoritma lain atau menambahkan fitur yang relevan untuk meningkatkan kinerja algoritma. Penelitian lanjutan juga dapat mempertimbangkan penggabungan beberapa fitur untuk melakukan klasfikasi penderita diabetes, serta menggunakan data lain yang lebih besar agar tidak terdapat data yang imbalance dan mampu menghasilkan keakuratan kinerja dari algoritma *Machine Learning*

DAFTAR PUSTAKA

Ainurrohma. (2021). Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka. *PRISMA, Prosiding Seminar Nasional Matematika*, *4*, 493–499. https://journal.unnes.ac.id/sju/index.php/prisma/

Angriani, S., & Baharuddin. (2020). Hubungan Tingkat Kecemasan Dengan Kadar Gula Darah Pada Penderita Diabetes Mellitus Tipe II Di Wilayah Kerja Puskesmas Batua Kota Makassar. *Jurnal Ilmiah Kesehatan Diagnosis*, *15*(2), 102–106.

Baiq Nurul Azmi, Arief Hermawan, & Donny Avianto. (2023). Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *4*(4), 281–290. https://doi.org/10.35746/jtim.v4i4.298

Fathurahman, H., Ariwikri, A., Pratama, G. A., Fikri, M. A. F. S., & Alrizki, M. F. (2023). Perbandingan Akurasi Metode Naive Bayes Classifier Dan Random Forest Menggunakan Reduksi Dimensi Linear Dicriminant Analysis (Lda) Untuk Diagnosis Penyakit Diabetes. *Jurnal Rekayasa Elektro Sriwijaya*, *4*(1), 24–31. https://doi.org/10.36706/jres.v4i1.58

Munir, A. S., Saputra, A. B., Aziz, A., & Barata, M. A. (2024). Perbandingan Akurasi Algoritma Naive Bayes dan Algoritma Decision Tree dalam Pengklasifikasian Penyakit Kanker Payudara. *Jurnal Ilmiah Informatika Global*, *15*(1), 23–29. https://doi.org/10.36982/jiig.v15i1.3578

Nur Azizah, A., Falach Asy'ari, M., Wisma Dwi Prastya, I., & Purwitasari, D. (2023). Easy Data Augmentation untuk Data yang Imbalance pada Konsultasi Kesehatan Daring. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(5), 1095–1104. https://doi.org/10.25126/jtiik.20231057082

Prasetyo, Y. A., Utami, E., & Yaqin, A. (2024). *Pengaruh Komposisi Split Data Terhadap Performa Akurasi Analisis Sentimen Algoritma Naïve Bayes dan SVM*. *6*(2), 382–390. https://doi.org/10.33650/jeecom.v4i2

Prastyo, P. H., Sumi, A. S., Dian, A. W., & Permanasari, A. E. (2020). Tweets Responding to the Indonesian Government's Handling of Covid-19: Sentiment Analysis Using SVM with Normalized Poly Kernel. *Journal of Information Systems Engineering and Business Intelligence*, *6*(2), 112. https://doi.org/10.20473/jisebi.6.2.112-122

Purnomo, A., Barata, M. A., Soeleman, M. A., & Alzami, F. (2020). Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease. *Journal of Physics: Conference Series*, *1511*(1). https://doi.org/10.1088/1742-6596/1511/1/012001

Ramon, E., Nazir, A., Novriyanto, N., Yusra, Y., & Oktavia, L. (2022). Klasifikasi Status Gizi Bayi Posyandu Kecamatan Bangun Purba Menggunakan Algoritma Support Vector

Machine (Svm). *Jurnal Sistem Informasi Dan Informatika (Simika)*, *5*(2), 143–150. https://doi.org/10.47080/simika.v5i2.2185

Sanjaya, U. P., Alawi, Z., Zayn, A. R., & Dirgantoro, G. P. (2023). Optimasi Convolutional Neural Network dengan Standard Deviasi untuk Klasifikasi Pneumonia pada Citra X-rays Paru. *Generation Journal*, *7*(3), 40–47. https://doi.org/10.29407/gj.v7i3.20183

Terbuka, P., Menurut, T. P. T., & Di, P. (2024). *Algoritma K-Means Untuk Mengelompokkan Tingkat*. *15*(2), 75–81.

Ucha Putri, S., Irawan, E., Rizky, F., Tunas Bangsa, S., -Indonesia Jln Sudirman Blok No, P. A., & Utara, S. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5. *Januari*, *2*(1), 39–46.

Yusnita, Y., Hi. A. Djafar, M., & Tuharea, R. (2021). Risiko Gejala Komplikasi Diabetes Mellitus Tipe II di UPTD Diabetes Center Kota Ternate. *Media Publikasi Promosi Kesehatan Indonesia (MPPKI)*, *4*(1), 60–73. https://doi.org/10.56338/mppki.v4i1.1391