

**ANALISIS KINERJA ALGORITMA MACHINE LEARNING
DENGAN METODE PREPROCESSING DATA ONE HOT
ENCODING DAN LABEL ENCODING DALAM
PREDIKSI PENYAKIT DIABETES**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 - Teknik Komputer



disusun oleh

WAHYU ENGGAR WICAKSONO

20.83.0495

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2024

**ANALISIS KINERJA ALGORITMA MACHINE LEARNING
DENGAN METODE PREPROCESSING DATA ONE HOT
ENCODING DAN LABEL ENCODING DALAM
PREDIKSI PENYAKIT DIABETES**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 - Teknik Komputer



disusun oleh

WAHYU ENGGAR WICAKSONO

20.83.0495

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2024

HALAMAN PERSETUJUAN

SKRIPSI

ANALISIS KINERJA ALGORITMA MACHINE LEARNING DENGAN
METODE PREPROCESSING DATA ONE HOT ENCODING DAN
LABEL ENCODING DALAM PREDIKSI PENYAKIT DIABETES

yang disusun dan diajukan oleh

Wahyu Enggar Wicaksono

20.83.0495

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 Juni 2024

Dosen Pembimbing,



Jeki Kuswanto, M.Kom.

NIK. 190302456

HALAMAN PENGESAHAN

SKRIPSI

ANALISIS KINERJA ALGORITMA MACHINE LEARNING DENGAN
METODE PREPROCESSING DATA ONE HOT ENCODING DAN
LABEL ENCODING DALAM PREDIKSI PENYAKIT DIABETES

yang disusun dan diajukan oleh

Wahyu Enggar Wicaksono

20.83.0495

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 Juni 2024

Susunan Dewan Penguji

Nama Penguji

Melwin Syafrizal, S.Kom., M.Eng
NIK. 190302105

Banu Santoso, S.T., M.Eng
NIK. 190302327

Jeki Kuswanto, M.Kom
NIK. : 190302456

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 Juni 2024

DEKAN FAKULTAS ILMU KOMPUTER



Hanif Al Fatta, S.Kom., M.Kom., Ph.D.
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Wahyu Enggar Wicaksono
NIM : 20.83.0495

Menyatakan bahwa Skripsi dengan judul berikut:

Analisis Kinerja Algoritma Machine Learning dengan Metode Preprocessing Data One Hot Encoding dan Label Encoding dalam Prediksi Penyakit Diabetes

Dosen Pembimbing : Jeki Kuswanto, M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 21 Juni 2024

Yang Menyatakan,



Wahyu Enggar Wicaksono

HALAMAN PERSEMBAHAN

Syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas rahmat dan hidayah-Nya yang mempermudah penulis dalam menyelesaikan skripsi ini. Penulis juga berterima kasih kepada semua pihak yang telah memberikan dukungan, inspirasi, dan motivasi, baik secara langsung maupun tidak langsung, sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Penulis mempersembahkan skripsi ini kepada:

1. Bapak Muhtarom dan Ibu Estuana Prajawati selaku orang tua penulis.
2. Bapak Jeki Kuswanto, M.Kom.
3. Semua teman - teman saya yang membantu dengan sepenuh hati.
4. Kerabat kerja yang memberikan saya cuti.

Semoga persembahan ini dapat mengungkapkan rasa terima kasih penulis kepada semua yang telah berperan dalam kesuksesan penyelesaian skripsi ini.

KATA PENGANTAR

Puji dan Syukur penulis panjatkan kehadiran Allah SWT, karena nikmat dan karunia-Nya penulis masih bisa menempuh pendidikan sampai saat ini, sekaligus dapat menyelesaikan skripsi yang berjudul “Analisis Kinerja Algoritma Machine Learning dengan Metode Preprocessing Data One Hot Encoding dan Label Encoding dalam Prediksi Penyakit Diabetes” ini.

Tak lupa penulis sampaikan rasa terima kasih untuk pihak-pihak yang telah berperan dalam penyusunan skripsi ini. Diantaranya adalah sebagai berikut:

1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Bapak Jeki Kuswanto, M.Kom. yang telah membimbing saya dalam proses menyusun skripsi ini.
3. Seluruh keluarga saya, terutama kedua orang tua yang telah membimbing dan mendidik saya hingga dewasa ini.
4. Teman - teman saya yang dengan setulus hati membantu penelitian ini hingga tuntas.
5. Kerabat kerja yang selalu mendukung saya untuk menyelesaikan pendidikan.

Masih banyak hal yang dirasa kurang dalam penyusunan proposal ini, maka dari itu penulis membuka hati dan diri sebesar-besarnya apabila ada kritik maupun saran guna melengkapi proposal ini.

Akhirnya semoga skripsi ini bisa bermanfaat bagi semua, serta menjadi acuan dan tuntunan penulis dalam membuat karya kedepan. Terima kasih.

Yogyakarta, 10 Juni 2024

Penulis

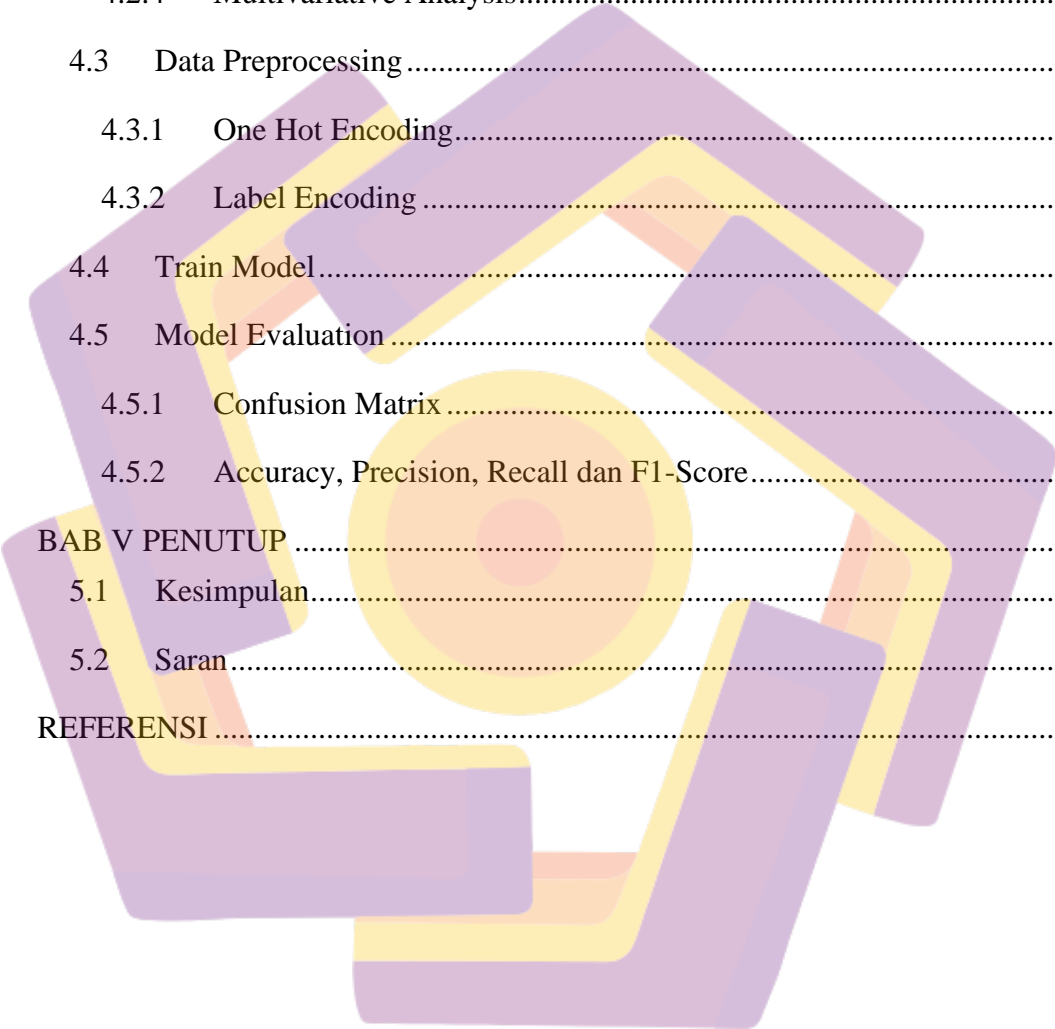


Wahyu Enggar Wicaksono

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	viii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMBANG DAN SINGKATAN	xv
DAFTAR ISTILAH.....	xvi
INTISARI	xviii
ABSTRACT.....	xix
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	4
1.6 Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA	5
2.1 Studi Literatur	5
2.2 Dasar Teori	15
2.2.1 Diabetes.....	15
2.2.2 Machine learning	16
2.2.3 Klasifikasi Machine Learning.....	18
2.2.4 Cara Kerja Machine Learning.....	18

2.2.5	Data Understanding	19
2.2.6	Data Preprocessing.....	22
2.2.7	Label Encoding	23
2.2.8	One Hot Encoding.....	24
2.2.9	Algoritma Machine Learning.....	24
2.2.10	Algoritma Naïve Bayes.....	25
2.2.11	Algoritma Support Vector Machine.....	26
2.2.12	Random Forest	27
2.2.13	Model Evaluation.....	28
2.2.14	Confusion Matrix	28
2.2.15	Accuracy	29
2.2.16	Precision.....	29
2.2.17	Recall	30
2.2.18	F-1 Score.....	30
BAB III	METODE PENELITIAN	31
3.1	Alur Penelitian.....	31
3.1.1	Data Collection	32
3.1.2	Data Understanding	33
3.1.3	Data Preprocessing.....	34
3.1.4	Train Model	36
3.1.5	Model Evaluation.....	37
3.2	Alat dan Bahan	38
3.1.1	Hardware.....	38
3.1.2	Software	38
BAB IV	HASIL DAN PEMBAHASAN	39
4.1	Data Collection.....	39



4.2	Data Understanding	41
4.2.1	Data Outlier	41
4.2.2	Visualisasi Fitur	43
4.2.3	Bivariate Analysis	51
4.2.4	Multivariate Analysis	56
4.3	Data Preprocessing	59
4.3.1	One Hot Encoding	59
4.3.2	Label Encoding	61
4.4	Train Model	65
4.5	Model Evaluation	68
4.5.1	Confusion Matrix	68
4.5.2	Accuracy, Precision, Recall dan F1-Score	70
BAB V PENUTUP		75
5.1	Kesimpulan	75
5.2	Saran	75
REFERENSI		76

DAFTAR TABEL

Tabel 2.1. Tabel Studi Literatur	5
Tabel 2.2. Tabel Studi Literatur	6
Tabel 2.3. Tabel Studi Literatur	6
Tabel 2.4. Tabel Studi Literatur	7
Tabel 2.5. Tabel Studi Literatur	8
Tabel 2.6. Tabel Studi Literatur	9
Tabel 2.7. Tabel Studi Literatur	9
Tabel 2.8. Tabel Studi Literatur	9
Tabel 2.9. Tabel Studi Literatur	11
Tabel 3.1. Tabel Hardware	38
Tabel 3.2. Tabel Software	38
Tabel 4.1. Hasil Evaluasi Model	71

DAFTAR GAMBAR

Gambar 2.1 Data Analysis	19
Gambar 2.2 Contoh Feature Analysis	20
Gambar 2.3 Contoh Bivariate Analysis	21
Gambar 2.4 Contoh Multivariate Analysis	22
Gambar 2.5 Label Encoding	23
Gambar 2.6 One Hot Encoding	24
Gambar 2.7 Algoritma Naive Bayes	25
Gambar 2.8 Algoritma Support Vector Machine	26
Gambar 2.9 Algoritma Random Forest	27
Gambar 2.10 Confusion Matrix	29
Gambar 3.1. Alur Penelitian	31
Gambar 3.2. Metode One-Hot Encoding	35
Gambar 3.3. Metode Label Encoding	36
Gambar 4.1. Dataset Diabetes from Kaggle	39
Gambar 4.2. Koneksi dengan Google Drive	39
Gambar 4.3. Import Library	40
Gambar 4.4. Variabel Dataset	40
Gambar 4.5. Dataset Diabetes	40
Gambar 4.6. Data Duplikat	41
Gambar 4.7. Data Unik	41
Gambar 4.8. Data Unik Fitur Gender	42
Gambar 4.9. Missing Value	43
Gambar 4.10. Import Library	43
Gambar 4.11. Code Visualisasi Fitur Age	44
Gambar 4.12. Visualisasi Fitur Age	44
Gambar 4.13. Code Visualisasi Gender	45
Gambar 4.14. Visualisasi Fitur Gender	45
Gambar 4.15. Code Visualisasi Fitur BMI	46
Gambar 4.16. Visualisasi Fitur BMI	46

Gambar 4.17. Code Visualisasi Variabel Biner	47
Gambar 4.18. Visualisasi Fitur Hypertension	48
Gambar 4.19. Visualisasi Fitur Heart Disease	48
Gambar 4.20. Visualisasi Label Diabetes Prediction	49
Gambar 4.21. Code Visualisasi Fitur Smoking History	49
Gambar 4.22. Visualisasi Smoking History	50
Gambar 4.23. Code Visualisasi BMI vs Diabetes	51
Gambar 4.24. Visualisasi BMI vs Diabetes	51
Gambar 4.25. Code Visualisasi Age vs Diabetes	52
Gambar 4.26. Visualisasi Age vs Diabetes	52
Gambar 4.27. Code Visualisasi Gender vs Diabetes	53
Gambar 4.28. Visualisasi Gender vs Diabetes	53
Gambar 4.29. Code Visualisasi HbA1c Level vs Diabetes	54
Gambar 4.30. Visualisasi HbA1c Level vs Diabetes	54
Gambar 4.31. Code Visualisasi Blood Glucose Level vs Diabetes	55
Gambar 4.32. Visualisasi Blood Glucose Level vs Diabetes	55
Gambar 4.33. Code Visualisasi Age vs BMI vs Diabetes	56
Gambar 4.34. Visualisasi Age vs BMI vs Diabetes	56
Gambar 4.35. Code Visualisasi Gender vs BMI vs Diabetes	57
Gambar 4.36. Visualisasi Gender vs BMI vs Diabetes	57
Gambar 4.37. Code Visualisasi Age vs Gender vs Diabetes	58
Gambar 4.38. Visualisasi Age vs Gender vs Diabetes	58
Gambar 4.39. Tipe Data Fitur	59
Gambar 4.40. Mengatur Fitur untuk One Hot Encoding	60
Gambar 4.41. Code Preprocessing One Hot Encoding	60
Gambar 4.42. Menggabungkan Data	60
Gambar 4.43. Hapus Fitur Gender dan Smoking History	61
Gambar 4.44. Hasil Dataset One Hot Encoding	61
Gambar 4.45. Data Unik	62
Gambar 4.46. Import Library Label Encoding	62
Gambar 4.47. Label Encoding Fitur Gender & Smoking History	63

Gambar 4.48. Code Visualisasi Hasil Label Encoding	63
Gambar 4.49. Visualisasi Fitur Smoking History	64
Gambar 4.50. Visualisasi Fitur Gender	64
Gambar 4.51. Import Library	65
Gambar 4.52. Split Data X dan Y (One Hot Encoding)	65
Gambar 4.53. Split Data X dan Y (Label Encoding)	66
Gambar 4.54. Split Data Train dan Test	66
Gambar 4.55. Algoritma SVM	67
Gambar 4.56. Algoritma Random Forest	67
Gambar 4.57. Algoritma Naïve Bayes	67
Gambar 4.58. Confusion Matrix	69
Gambar 4.59. Prediksi Data Testing	70
Gambar 4.60. Perbandingan Hasil Accuracy	71
Gambar 4.61. Perbandingan Hasil Precision	72
Gambar 4.62. Perbandingan Hasil Recall	73
Gambar 4.63. Perbandingan Hasil F1-Score	73

DAFTAR LAMBANG DAN SINGKATAN

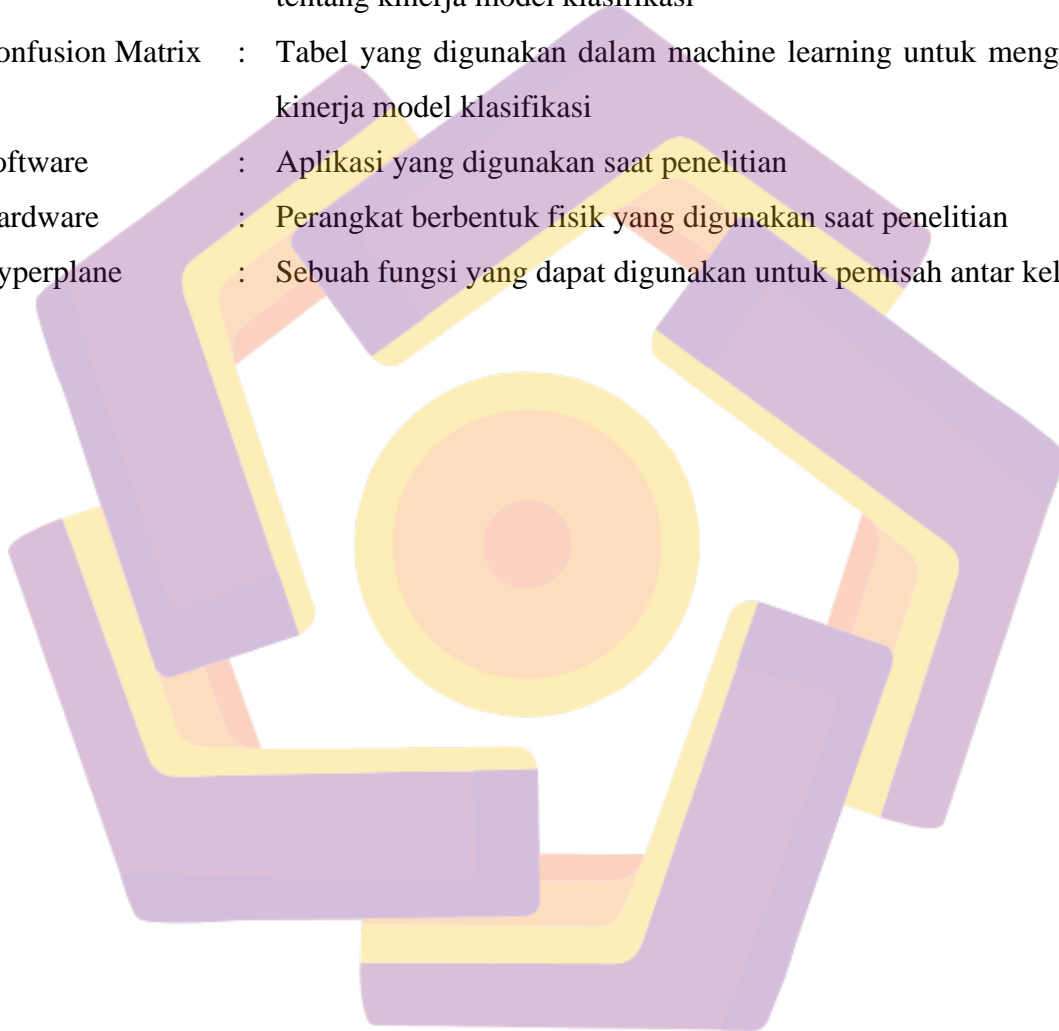


AI	: <i>Artificial Intelligence</i>
ML	: <i>Machine Learning</i>
IDF	: <i>International Diabetes Federation</i>
SVM	: <i>Support Vector Machines</i>
NB	: <i>Naive Bayes</i>
DT	: <i>Decision Trees</i>
KKN	: <i>K Means Clustering</i>
DNN	: <i>Deep Neural Networks</i>
AUC	: <i>Area Under the Curve</i>
ROC	: <i>Receiver Operating Characteristic</i>
BMI	: <i>Body Mass Index</i>
TP	: <i>True Positive</i>
TN	: <i>True Negative</i>
FP	: <i>False Positive</i>
FN	: <i>False Negative</i>

DAFTAR ISTILAH

Insulin	:	Hormon pankreas untuk mengatur kadar gula dalam darah
Dataset	:	Sekumpulan data yang digunakan untuk melatih model
Model	:	Representasi matematis dari hubungan antara variabel input dan output yang dipelajari dari data
Machine Learning	:	Teknologi yang memungkinkan mesin belajar dari data tanpa diprogram secara eksplisit
Artificial Intelligence	:	Teknologi yang memungkinkan mesin untuk meniru kecerdasan manusia
Parameter Evaluasi	:	Metrik yang digunakan untuk menilai kinerja atau efektivitas suatu sistem, model, atau proses dalam mencapai tujuan tertentu
Algoritma	:	Langkah - langkah yang perlu komputer lakukan untuk melakukan tugas tertentu atau menyelesaikan suatu masalah
Preprocessing Data	:	Proses mempersiapkan dan membersihkan data agar sesuai untuk analisis atau penggunaan dalam pemodelan
One Hot Encoding	:	Teknik mengubah data kategorikal menjadi bilangan biner
Label Encoding	:	Metode mengubah data label teks menjadi format numerik
Testing	:	Proses menguji kinerja model pada data yang tidak digunakan selama pelatihan untuk mengevaluasi seberapa baik model tersebut dapat melakukan prediksi pada data baru
Training	:	Proses di mana model belajar dari data yang diberikan dengan mengidentifikasi pola dan hubungan di dalamnya untuk membuat prediksi atau mengambil keputusan
Accuracy	:	Metrik yang mengukur seberapa tepat model dalam membuat prediksi yang benar dibandingkan dengan total jumlah prediksi yang dilakukan
Precision	:	Metrik yang mengukur proporsi dari prediksi positif yang benar dibandingkan dengan total jumlah prediksi positif yang dibuat oleh model

- Recall : Metrik yang mengukur proporsi dari semua kasus positif yang berhasil diidentifikasi oleh model, dibandingkan dengan total jumlah kasus positif yang sebenarnya
- F-1 Score : Metrik tunggal yang menggabungkan precision dan recall dalam sebuah nilai tunggal untuk memberikan gambaran keseluruhan tentang kinerja model klasifikasi
- Confusion Matrix : Tabel yang digunakan dalam machine learning untuk mengevaluasi kinerja model klasifikasi
- Software : Aplikasi yang digunakan saat penelitian
- Hardware : Perangkat berbentuk fisik yang digunakan saat penelitian
- Hyperplane : Sebuah fungsi yang dapat digunakan untuk pemisah antar kelas



INTISARI

Diabetes adalah penyakit yang terjadi karena kadar gula dalam darah terlalu tinggi. Gula yang seharusnya diubah menjadi energi tidak dapat diserap oleh tubuh, sehingga menumpuk dalam sel. Menurut International Diabetes Federation (IDF), pada tahun 2021 terdapat 537 juta penderita diabetes di seluruh dunia, termasuk 19 juta di Indonesia, dan jumlah ini terus bertambah setiap tahun. Saat ini, teknologi *machine learning* bisa digunakan untuk memprediksi risiko diabetes. Namun, hasil prediksi ini sangat bergantung pada *dataset*, metode *preprocessing* data, dan algoritma yang digunakan. Penelitian ini bertujuan untuk menganalisis kinerja tiga algoritma *machine learning* *Support Vector Machine (SVM)*, *Random Forest* dan *Naive Bayes* dalam memprediksi diabetes dengan menggunakan dua teknik *preprocessing* data: *One Hot Encoding* dan *Label Encoding*. *Dataset* yang digunakan adalah *Diabetes Prediction Dataset*. Penelitian ini membandingkan bagaimana setiap algoritma beradaptasi dengan teknik *encoding* yang berbeda dan mengevaluasi kinerjanya menggunakan metrik *accuracy*, *precision*, *recall* dan *f1-score* melalui metodologi pengujian yang terstruktur. Hasil penelitian menunjukkan bahwa di antara ketiga algoritma tersebut, *Random Forest* dengan *preprocessing* *One Hot Encoding* dan *Label Encoding* memiliki kinerja terbaik. Dalam hal *precision*, *One Hot Encoding* lebih unggul dengan selisih 0,22% dibandingkan *Label Encoding*. Sebaliknya, dalam hal *recall*, *Label Encoding* lebih baik dengan selisih 0,12% dibandingkan *One Hot Encoding*. Untuk parameter *accuracy* dan *f1-score*, keduanya memberikan hasil yang sama, yaitu *accuracy* sebesar 96,68% dan *f1-score* sebesar 78,07%.

Kata kunci: *machine learning, preprocessing, one hot encoding, label encoding*

ABSTRACT

Diabetes is a disease that occurs because blood sugar levels are too high. Sugar that should be converted into energy cannot be absorbed by the body, so it accumulates in the cells. According to the International Diabetes Federation (IDF), in 2021 there will be 537 million people with diabetes worldwide, including 19 million in Indonesia, and this number continues to grow every year. Currently, Machine Learning technology can be used to predict diabetes risk. However, the results of these predictions are highly dependent on the dataset, data preprocessing methods, and algorithms used. This study aims to analyze the performance of three machine learning algorithms Support Vector Machine (SVM), Random Forest and Naive Bayes in predicting diabetes using two data preprocessing techniques: One Hot Encoding and Label Encoding. The dataset used is the Diabetes Prediction Dataset. This research compares how each algorithm adapts to different encoding techniques and evaluates its performance using accuracy, precision, recall and f1-score metrics through a structured testing methodology. The results show that among the three algorithms, Random Forest with One Hot Encoding and Label Encoding preprocessing has the best performance. In terms of precision, One Hot Encoding is superior with a difference of 0.22% compared to Label Encoding. Conversely, in terms of recall, Label Encoding is better with a difference of 0.12% compared to One Hot Encoding. For accuracy and f1-score parameters, both provide the same results, namely accuracy of 96.68% and f1-score of 78.07%.

Keyword: machine learning, preprocessing, one hot encoding, label encoding