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

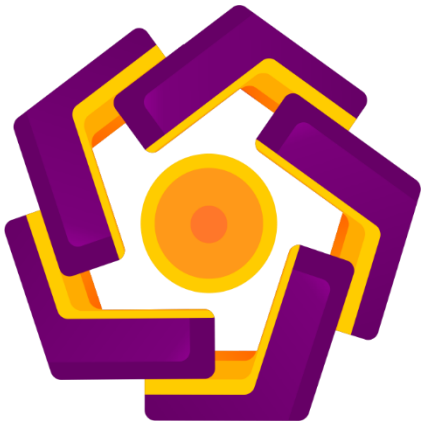
2024

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	Tanda Tangan
Melwin Syafrizal, S.Kom., M.Eng **NIK. 190302105**	
Banu Santoso, S.T., M.Eng **NIK. 190302327**	
Jeki Kuswanto, M.Kom **NIK. : 190302456**	

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 kehadirat 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:

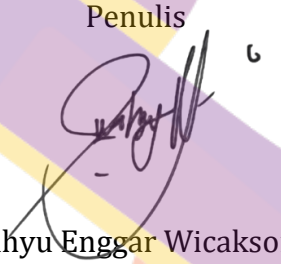
1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Bapak Jeki Kuswanto, M.Kom. yang telah membimbing saya dalam proses menyusun skripsi ini.
3. Seluruh keluarga saya, terutama kedua orang tua yang telah membimbing dan mendidik saya hingga sedewasa ini.
4. Teman - teman 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

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMBANG DAN SINGKATAN

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

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*