

**OPTIMASI ALGORITMA DATA MINING MENGGUNAKAN
BACKWARD ELIMINATION UNTUK KLASIFIKASI PENYAKIT
DIABETES**

SKRIPSI



disusun oleh

Muhammad Abid Wiratama

18.11.1938

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2022**

**OPTIMASI ALGORITMA DATA MINING MENGGUNAKAN
BACKWARD ELIMINATION UNTUK KLASIFIKASI PENYAKIT
DIABETES**

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika



disusun oleh

Muhammad Abid Wiratama

18.11.1938

**PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2022**

PERSETUJUAN

SKRIPSI

**OPTIMASI ALGORITMA DATA MINING MENGGUNAKAN
BACKWARD ELIMINATION UNTUK KLASIFIKASI PENYAKIT
DIABETES**

yang dipersiapkan dan disusun oleh

Muhammad Abid Wiratama

18.11.1938

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 Januari 2022

Dosen Pembimbing,

Windha Mega Pradnya D, M.Kom

NIK. 190302185

PENGESAHAN
SKRIPSI
OPTIMASI ALGORITMA DATA MINING MENGGUNAKAN
BACKWARD ELIMINATION UNTUK KLASIFIKASI PENYAKIT
DIABETES

yang dipersiapkan dan disusun oleh

Muhammad Abid Wiratama

18.11.1938

telah dipertahankan di depan Dewan Penguji
pada tanggal 17 Februari 2022

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Arifiyanto Hadinegoro, S.Kom, MT
NIK. 190302289

Erni Seniwati, S.Kom., M.Cs.
NIK. 190302231

Windha Mega Pradnya D, M.Kom
NIK. 190302185

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 17 Februari 2022

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta, M.Kom
NIK.190302096

PERNYATAAN KEASLIAN SKRIPSI

PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertandatangan di bawah ini menyatakan bahwa, skripsi ini karya saya sendiri (ASLI), dan isi dalam skripsi ini tidak terdapat karya yang pernah diajukan oleh orang lain untuk memperoleh gelar akademis di suatu institusi pendidikan tinggi manapun, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis dan/atau diterbitkan oleh orang lain, kecuali yang secara tertulis diajau dalam naskah ini dan disebutkan dalam daftar pustaka.

Segala sesuatu yang terkait dengan naskah dan karya yang telah dibuat adalah menjadi tanggungjawab saya pribadi.

Yogyakarta, 7 Februari 2022



Muhammad Abid Wiratama

NIM. 18.11.1938

MOTTO

“Angin tidak berhembus untuk menggoyangkan pepohonan, melainkan menguji kekuatan akarnya.”

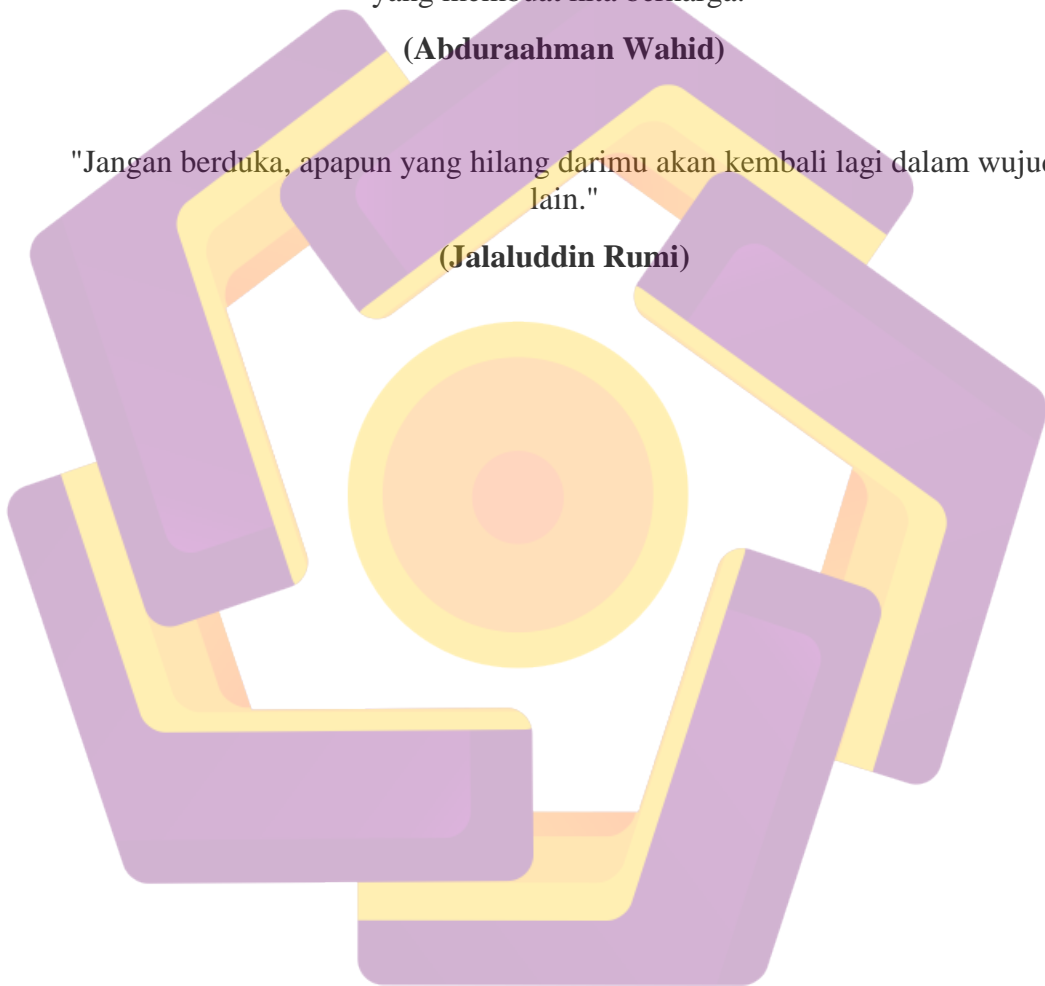
(Ali Bin Abi Thalib)

"Menyesali nasib tidak akan mengubah keadaan. Terus berkarya dan bekerjalah yang membuat kita berharga."

(Abduraahman Wahid)

"Jangan berduka, apapun yang hilang darimu akan kembali lagi dalam wujud lain."

(Jalaluddin Rumi)



PERSEMBAHAN

Persembahan tugas akhir dan rasa terimakasih kepada :

1. Allah SWT Pencipta alam semesta yang memberiku hidup yang berkah dan rizkinya.
2. Kedua orang tua yaitu bapak dan ibu tercinta yang selalu memberikan dukungannya dan pengorbanannya yang tak terkira harganya, doa serta kasih sayangnya tak akan pernah ananda lupakan sepanjang hidup.
3. Untuk seluruh keluargaku dan saudara dimanapun berada yang secara tidak langsung memberikan dorongan dan semangat yang tak pernah putus sehingga tugas akhir ini berjalan lancar.
4. Teruntuk sahabat dan kawan seperjuangan yang selalu memberikan semangat semoga kita sukses bersama di masa depan kelak.

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan kepada kehadiran Allah SWT, karena berkat rahmat dan karunia-Nya, penulis dapat menyelesaikan tugas akhir atau skripsi yang berjudul “ Optimasi Algoritma Data Mining Menggunakan Backward Elimination Untuk Klasifikasi Penyakit Diabetes”.

Pengerjakan Skripsi ini dapat diselesaikan tidak lepas dari bantuan banyak pihak baik dukungan moral maupun dukungan lainnya, untuk itu sudah sepantasnya saya mengucapkan terimakasih kepada mereka yang telah berjasa membantu penulis dari awal penelitian dilakukan sampai ketahap ini, semoga mereka mendapatkan berkah dari pengerjakan skripsi ini dan senantiasa diberikan rahmat dan hidayah dari Allah SWT.

Namun, penulis menyadari bahwa penelitian dalam skripsi ini masih belum sempurna dan masih memiliki banyak kekurangan, baik dari segi tata bahasa maupun isi dari penelitian itu sendiri. Oleh karena itu, dengan segala kekurangan dan kerendahan hati, saya sangat mengharapkan kritik dan saran dari pemaca sekalian agar nantinya skripsi ini menjadi lebih baik.

Akhirnya, penulis berharap agar karya ilmiah ini dapat memberikan dampak positif baik untuk masyarakat maupun pihak yang terkait dan semoga dapat memajukan ilmu pengetahuan di bidang ini.

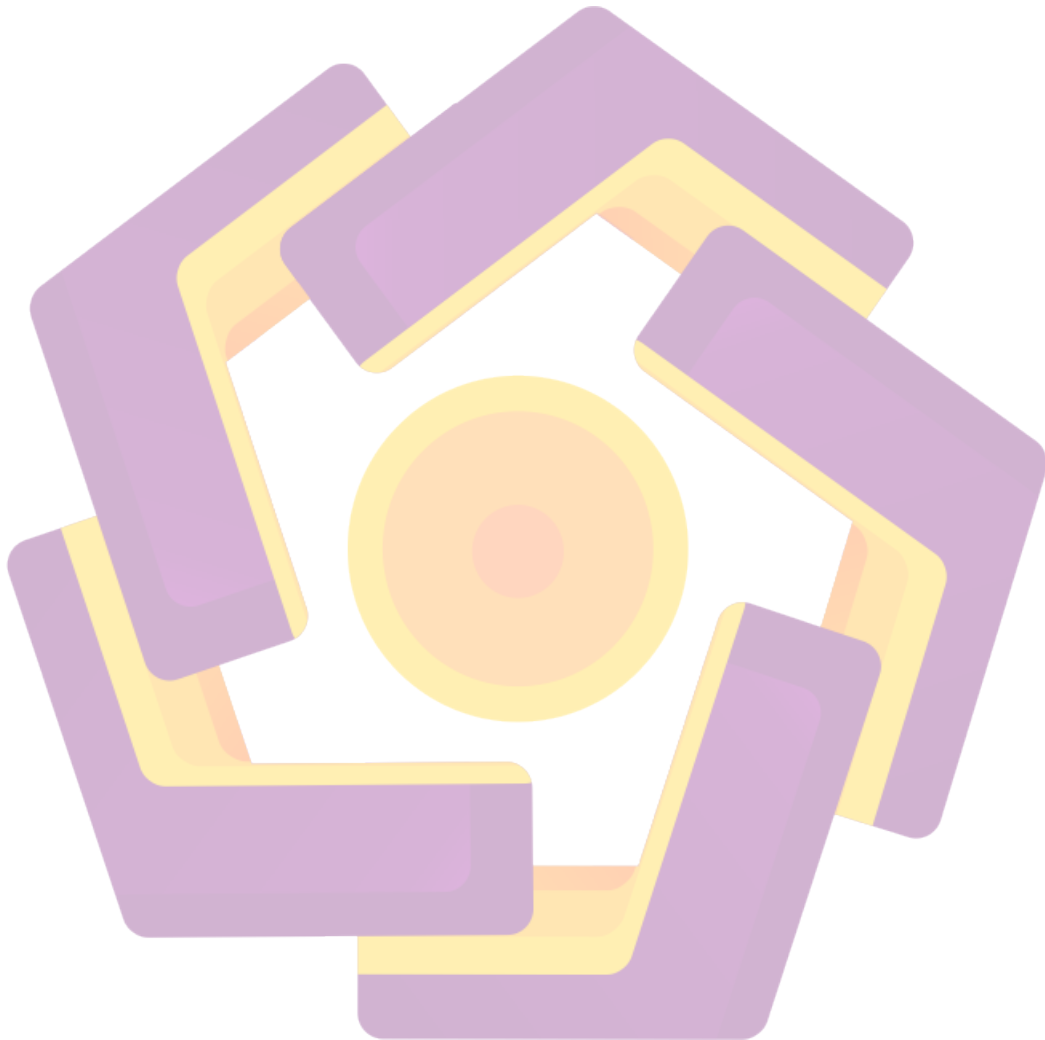
DAFTAR ISI

PERSETUJUAN	I
PENGESAHAN	II
PERNYATAAN KEASLIAN SKRIPSI.....	III
MOTTO.....	IV
PERSEMBAHAN.....	V
KATA PENGANTAR	VI
DAFTAR ISI.....	VII
DAFTAR TABEL.....	XI
DAFTAR GAMBAR	XII
INTISARI.....	XIII
<i>ABSTRACT</i>	XIV
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah	6
1.3 Batasan Masalah.....	6
1.4 Tujuan Penelitian.....	7
1.5 Manfaat Penelitian.....	7
1.5.1 Bagi Objek.....	8
1.5.2 Bagi Peneliti.....	8
1.5.3 Bagi Universitas Amikom Yogyakarta	8
1.5.4 Bagi Masyarakat.....	9
1.5.5 Bagi Peneliti Selanjutnya	9
1.6 Metodologi Penelitian	9
1.6.1 Metode Pengumpulan Data.....	9

1.6.2	Metode Analisis	10
1.7	Sistematika Penulisan.....	12
BAB II LANDASAN TEORI		14
2.1	Tinjauan Pustaka	14
2.2	Dasar Teori.....	22
2.2.1	Diabetes.....	22
2.2.2	Data Mining	23
2.2.3	Klasifikasi	28
2.2.4	Algoritma C4.5.....	29
2.2.5	Algoritma Naïve Bayes.....	30
2.2.6	Algoritma K-Nearest Neighbor.....	32
2.2.7	Feature Selection.....	32
2.2.8	Backward Elimination.....	33
2.2.9	K-fold Cross Validation	36
2.2.10	Uji Paired T-Test.....	36
2.2.11	Google Interactive Notebook (Google Colab)	38
BAB III METODOLOGI PENELITIAN.....		40
3.1	Alat dan Bahan Penelitian	40
3.1.1	Alat.....	40
3.1.2	Bahan.....	41
3.2	Alur Penelitian.....	41
3.2.1	Business Understanding	42
3.2.2	Data Understanding.....	42
3.2.3	Data Preparation.....	45
3.2.4	Modelling	48

3.2.5	Evaluation	52
3.2.6	Deployment.....	53
BAB IV HASIL DAN PEMBAHASAN		54
3.3	Business Understanding	54
3.4	Data Understanding	54
3.5	Data Preparation	58
3.5.1	Data Cleaning.....	58
3.5.2	Tranform Data.....	59
3.6	Modelling	60
3.6.1	Model KNN.....	61
3.6.2	Model Naïve Bayes.....	61
3.6.3	Model C4.5.....	61
3.6.4	Model KNN + Backward Elimination	62
3.6.5	Model Naïve Bayes + Backward Elimination.....	62
3.6.6	Model C4.5 + Backward Elimination	63
3.6.7	Uji Paired T-test.....	64
3.7	Evaluation.....	64
3.7.1	Evaluasi Model KNN.....	64
3.7.2	Evaluasi Model Naïve Bayes.....	66
3.7.3	Evaluasi Model C4.5.....	67
3.7.4	Evaluasi Model KNN + Backward Elimination	69
3.7.5	Evaluasi Model Naïve Bayes + Backward Elimination.....	72
3.7.6	Evaluasi Model C4.5 + Backward Elimination.....	75
3.7.7	Uji Paired T-Test.....	78
3.8	Deployment	79

BAB V KESIMPULAN DAN SARAN.....	81
4.1 Kesimpulan.....	81
4.2 Saran.....	82
DAFTAR PUSTAKA	83



DAFTAR TABEL

Tabel 3.1 Spesifikasi Hardware.....	40
Tabel 3.2 <i>Sampel Dataset</i>	42
Tabel 3.3 <i>Penjabaran Atribut</i>	43
Tabel 3.4 <i>Contoh Data Dengan Missing Value</i>	44
Tabel 3.5 Sampel Yang Memiliki Missing Value	46
Tabel 3.6 <i>Sampel Yang Sudah Dilakukan Data Cleaning</i>	46
Tabel 3.7 Sampel Yang Sudah Dilakukan Transform Data	47
Tabel 4.1 <i>Sampel Dataset</i>	54
Tabel 4.2 <i>Penjabaran Atribut</i>	56
Tabel 4.3 Check Missing Value	58
Tabel 4.4 Check Missing Value Setelah Data Cleaning	59
Tabel 4.5 Validasi Akurasi Model KNN Dengan 10-fold Validation.....	64
Tabel 4.6 Validasi Akurasi Model Naive Bayes Dengan 10-fold Cross Validation	66
Tabel 4.7 Akurasi Validasi Model C4.5 Dengan 10-fold Cross Validation.....	67
Tabel 4.8 Tabel hasil model akurasi dan auc	79

DAFTAR GAMBAR

Gambar 1.1 Jumlah Kasus Diabetes Di Indonesia	2
Gambar 2.1 Proses Data Mining dalam knowledge discovery.....	23
Gambar 2.2 Fase Data Mining dalam CRISP-DM.....	25
Gambar 3.1 Flowchart Alur Penelitian.....	41
Gambar 3.2 Langkah Data Preparation	45
Gambar 3.3 Langkah Modelling.....	48
Gambar 3.4 Flowchart model tanpa optimasi.....	49
Gambar 3.5 Flowchart Model dengan Optimasi	50
Gambar 4.1 Tranform Data	60
Gambar 4.2 Plotting Validasi Akurasi Model KNN 10-fold Cross Validation...	65
Gambar 4.3 Kurva ROC dan Nilai AUC Model KNN.....	65
Gambar 4.4 Plotting Validasi Akurasi Model Naive Bayes	66
Gambar 4.5 Kurva ROC dan Nilai AUC Model Naive Bayes	67
Gambar 4.6 Plotting Validasi Akurasi Model C4.5.....	67
Gambar 4.7 Kurva ROC dan Nilai AUC Model C4.5.....	68
Gambar 4.8 Tabel Validasi Akurasi Model KNN + Backward Elimination.....	69
Gambar 4.9 Plotting Validasi Akurasi Model KNN + Backward Elimination ...	70
Gambar 4.10 Validasi Nilai AUC Model KNN + Backward Elimination	71
Gambar 4.11 Tabel Validasi Akurasi Model Naive Bayes + Backward Elimination	72
Gambar 4.12 Plotting Validasi Akurasi Model Naive Bayes + Backward Elimination.....	73
Gambar 4.13 Validasi Nilai AUC Model Naive Bayes + Backward Elimination	74
Gambar 4.14 Tabel Validasi Akurasi Model C4.5 + Backward Elimination.....	75
Gambar 4.15 Plotting Validasi Akurasi Model C4.5 + Backward Elimination ..	76
Gambar 4.16 Validasi Nilai AUC Model C4.5 + Backward Elimination	77
Gambar 4.17 Hasil Uji Paired T-Test.....	78

INTISARI

Diabetes Mellitus (DM) adalah penyakit metabolik kronis yang ditandai dengan peningkatan kadar gula darah, yang dengan seiring berjalannya waktu akan menyebabkan kerusakan pada organ tubuh lainnya. Menurut situs resmi World Health Organization (WHO) sekitar 422 juta orang di seluruh dunia menderita diabetes. Di Indonesia sendiri menurut situs databox, Indonesia menempati urutan ke-5 kasus diabetes terbanyak di dunia. Keterlambatan dalam diagnosis penyakit diabetes adalah satu penyebab terjadinya lonjakan jumlah kematian maka dari itu tindakan awal yang harus dilakukan adalah deteksi dini..

Dari banyaknya kasus tersebut maka dihasilkan data pasien diabetes yang dapat diolah sehingga penelitian ini dibuat untuk mengetahui algoritma klasifikasi terbaik yang mempunyai nilai akurasi dan nilai AUC tertinggi untuk melakukan deteksi dini penyakit diabetes. Algoritma yang akan teliti adalah algoritma KNN, Naïve Bayes, dan C4.5. Algoritma akan dilakukan optimasi menggunakan metode backward elimination. Backward Elimination dilakukan untuk mengoptimalkan hasil akurasi dari kinerja algoritma. Penelitian ini akan dilakukan menggunakan metode CRISP-DM yaitu salah satu framework untuk melakukan data mining.

Setelah melalui tahapan-tahapan metode CRISP-DM, hasil penelitian adalah algoritma KNN sebelum dioptimasi mendapatkan akurasi 0,928 dan nilai auc 0,942, sedangkan setelah dioptimasi mendapatkan akurasi 0,976 dan nilai auc 0,973, algoritma Naïve Bayes sebelum dioptimasi mendapatkan akurasi 0,880 dan nilai auc 0,912, sedangkan setelah dioptimasi mendapatkan akurasi 0,894 dan nilai auc 0,958, algoritma C4.5 sebelum dioptimasi mendapatkan akurasi 0,967 dan nilai auc 0,956, sedangkan setelah dioptimasi mendapatkan akurasi 0,975 dan nilai auc 0,988. Sehingga kesimpulan dari penelitian ini adalah bahwasanya algoritma terbaik dari segi akurasi adalah algoritma KNN yang sudah dioptimasi dengan nilai akurasi 0,976 dan dari segi nilai auc yang dihasilkan adalah algoritma C4.5 yang sudah dioptimasi dengan nilai auc 0,988.

Kata Kunci : *Diabetes Mellitus, Klasifikasi, Backward Elimination, KNN, Naïve Bayes, C4.5*

ABSTRACT

Diabetes Mellitus (DM) is a chronic metabolic disease characterized by increased blood sugar levels, which over time will cause damage to other organs of the body. According to the official website of the World Health Organization (WHO), about 422 million people worldwide suffer from diabetes. In Indonesia, according to the databox site, Indonesia ranks the 5th most cases of diabetes in the world. The delay in diagnosing diabetes is one cause of the spike in the number of deaths, therefore the initial action that must be taken is early detection.

From these many cases, data from diabetes patients is generated that can be processed so that this study was made to determine the best classification algorithm that has the highest accuracy and AUC values for early detection of diabetes. The algorithms that will be studied are the KNN, Naïve Bayes, and C4.5 algorithms. The algorithm will be optimized using the backward elimination method. Backward Elimination is done to optimize the accuracy of the algorithm's performance. This research will be conducted using the CRISP-DM method, which is one of the frameworks for performing data mining.

After going through the stages of the CRISP-DM method, the results of the research are that the KNN algorithm before being optimized got an accuracy of 0.928 and an auc value of 0.942, while after being optimized it got an accuracy of 0.976 and an auc value of 0.973, the Naïve Bayes algorithm before optimization got an accuracy of 0.880 and an auc value of 0.912, while after being optimized it gets an accuracy of 0.894 and an auc value of 0.958, the C4.5 algorithm before being optimized gets an accuracy of 0.967 and an auc value of 0.956, while after being optimized it gets an accuracy of 0.975 and an auc value of 0.988. So the conclusion from this research is that the best algorithm in terms of accuracy is the KNN algorithm which has been optimized with an accuracy value of 0.976 and in terms of the resulting AUC value is the C4.5 algorithm which has been optimized with an AUC value of 0.988.

Keywords : *Diabetes Mellitus, Classification, Backward Elimination, KNN, Naïve Bayes, C4.5*