PENERAPAN DATA MINING UNTUK PREDIKSI AWAL KEMUNGKINAN TERINDIKASI DIABETES

Erfan Karyadiputra[1]*, Agus Setiawan[2]
[1]Program Studi Sistem Informasi
Fakultas Teknologi Informasi Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin
Jl. Adhyaksa No. 2, Banjarmasin, Kalimantan Selatan, Indonesia. 70104
*E-mail: erfantsy@gmail.com
[2]Program Studi Teknik Informatika
Fakultas Teknologi Informasi Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin
Jl. Adhyaksa No. 2, Banjarmasin, Kalimantan Selatan, Indonesia. 70104

Abstrak: Diabetes merupakan salah satu penyakit kronis yang memiliki ciri khas berupa tingginya kadar gula (glukosa) darah. Glukosa adalah sumber energi paling utama bagi sel-sel tubuh manusia, namun glukosa yang tertumpuk di dalam darah dapat mengakibatkan berbagai gangguan pada organ tubuh jika tidak dikontrol dan menimbulkan berbagai komplikasi penyakit lain yang dapat membahayakan penderitanya. Deteksi dini diabetes diperlukan karena adanya fase asimtomatik yang cukup lama, fase asimtomatik adalah kondisi penyakit yang sudah positif diderita tetapi tidak menimbulkan gejala klinis pada penderita. Tujuan penelitian ini adalah untuk memprediksi kemungkinan awal seseorang terindikasi penyakit diabetes berdasarkan dataset diabetes menggunakan teknik data mining yaitu metode algoritma *Decision Tree C4.5, Naive Bayes dan K-Nearest Neighbors.* Hasil pengujian *Cross Validation* dari ketiga algortima kemudian dibandingkan dengan pengukuran *performance menggunakan Confusion Matrix, Compare ROC dan Paired T-Test* sehingga didapatkan metode algortima terbaik. Adapun hasil dari penelitian ini menunjukkan bahwa algoritma Decision Tree C4.5 menjadi algoritma terbaik berdasarkan hasil *performance* akurasi prediksi sebesar 96,35% dengan nilai AUC sebesar 0,949 sehingga termasuk ke dalam kategori *excellent classification* serta hasil uji beda T-Test yang dominan jika dibandingkan dengan algoritma lainnya. Oleh sebab itu algortima *Decision Tree C4.5* lebih akurat dalam memprediksi awal kemungkinan seseorang terindikasi penyakit diabetes.

Kata Kunci : asimtomatik, data mining, *decision tree*, deteksi dini, diabetes, *naive bayes*, *k-nearest neighbors*

Abstract: Diabetes is one of the chronic diseases that have a characteristic in the form of high levels of blood sugar (glucose). Glucose is the main source of energy for the cells of the human body, but glucose accumulated in the blood can cause various disorders of the body's organs if not controlled and cause various complications of other diseases that can harm the sufferer. Early detection of diabetes is necessary because of the long asymptomatic phase, the asymptomatic phase is a condition of the disease that has been positively suffered but does not cause clinical symptoms in sufferers. The purpose of this study was to predict the initial likelihood of a person being indicated by

diabetes based on the diabetes dataset using data mining techniques, namely the Decision Tree C4.5 algorithm method, Naive Bayes and K-Nearest Neighbors. Cross Validation test results from the three algoritma were then compared with performance measurements using Confusion Matrix, Compare ROC and Paired T-Test so that the best algorithm method was obtained. The results of this study showed that the Decision Tree C4.5 algorithm became the best algorithm based on the results of predictive accuracy performance of 96.35% with an AUC value of 0.949 so that it belongs to the category of excellent classification and the results of different T-Test tests are dominant when compared to other algorithms. Therefore, the decision tree C4.5 algorithm is more accurate in predicting the beginning of the possibility of someone being indicated by diabetes.

Keywords: asymptomatic, data mining, decision tree, early detection, diabetes, naive bayes, k-nearest neighbors

PENDAHULUAN

Diabetes merupakan salah satu penyakit kronis yang memiliki ciri khas berupa tingginya kadar gula (glukosa) darah (Handayanna, 2012). Glukosa merupakan sumber utama energi terutama bagi kebutuhan sel-sel pada tubuh manusia, namun glukosa yang banyak tertumpuk di dalam darah dapat mengakibatkan berbagai gangguan fungsi organ tubuh jika tidak dikontrol dan dapat menimbulkan berbagai komplikasi yang membahayakan nyawa bagi penderitanya (Kementerian Kesehatan RI, 2021). Berdasarkan data International Diabetes Federation (IDF) pada tahun 2021, sekitar 537 juta orang dewasa (20-79 tahun) hidup dengan diabetes dan Indonesia berstatus waspada diabetes karena menduduki peringkat ke-7 dari 10 negara dengan jumlah pasien diabetes tertinggi di dunia. Prevalensi pasien penderita penyakit diabetes di Indonesia sudah mencapai 6,2 persen, yang berarti ada sekitar 10,8 juta orang penderita diabetes per tahun 2020 (International Diabetes Federation, 2021).

Terdapat beberapa cara diagnosa diabetes yang bisa dilakukan yaitu dengan mengukur kadar glukosa yang terkandung dalam darah, apabila kadar glukosa dalam darah konsentrasinya melebihi batas normal maka orang tersebut dikategorikan menderita penyakit diabetes (Detikhealth, 2019). Namun deteksi dini diabetes diperlukan karena adanya fase asimtomatik yang cukup lama, fase asimtomatik adalah kondisi penyakit yang sudah positif diderita tetapi tidak menimbulkan gejala klinis pada penderita (Eldridge, 2022). Diagnosis dini diabetes hanya mungkin dengan penilaian yang tepat dari kedua gejala tanda umum dan kurang umum, yang dapat ditemukan dalam fase yang berbeda dari inisiasi penyakit hingga diagnosis (Islam et al., 2020).

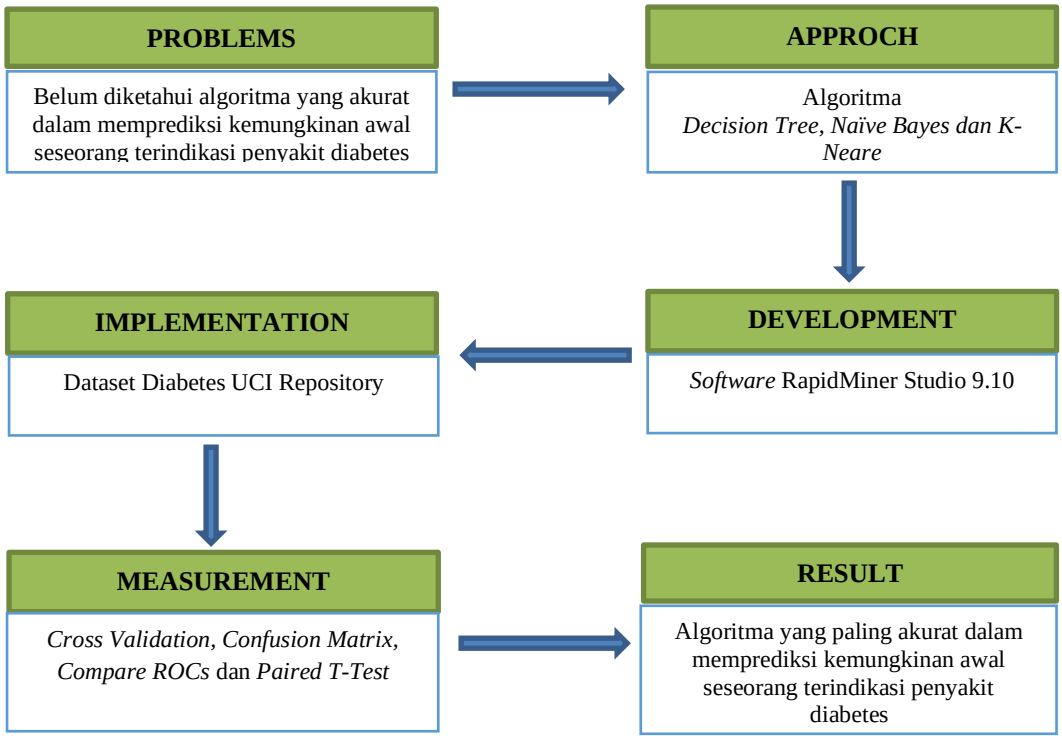
Data mining diterapkan untuk mendapatkan pola informasi baru berupa model dari suatu database (Anggraeni & Ramadhani, 2018). Peranan data mining dalam bidang kesehatan memiliki potensi besar terutama untuk menemukan pola-pola yang tersembunyi dalam suatu dataset rekam medis yang kemudian pola-pola tersebut dimanfaatkan untuk mendiagnosa awal penyakit secara klinis (Edy, 2012). Beberapa metode algoritma dalam data mining yang sering digunakan untuk memprediksi seperti *Decision Tree C4.5, Naive Bayes* dan *K-Nearest Neighbors.* Akurasi model *Decision Tree* C4.5 sangat tergantung pada atribut pembentuknya, oleh karena itu pengolahan awal data sangat penting untuk menghasilkan model sederhana yang tingkat akurasinya tinggi

(Lasut, 2012). Kelebihan menggunakan model klasifikasi *Decision Tree C4.5* yaitu hasil dari model pohon keputusan sederhana dan mudah dimengerti (Suradiradja, 2012). Adapun algoritma *naive bayes* mempunyai tingkat akurasi dan kecepatan tinggi terutama saat diimplementasikan pada data besar berdasarkan probabilitas keanggotaan suatu kelas (Rumini & Nasruddin, 2021). Metode algoritma *K-Nearest Neighbor* melakukan klasifikasi berdasarkan pencocokan dari nilai sejumlah fitur tetangga terdekatnya dengan menghitung kedekatan kasus lama dengan kasus baru (Mustafa & Simpen, 2019). Berdasarkan penjelasan dari uraian latar belakang permasalahan, maka penelitian ini bertujuan untuk mendapatkan metode algortima paling akurat dalam memprediksi awal kemungkinan seseorang terindikasi penyakit diabetes. Hasil prediksi akan membantu dalam menentukan hipotesa awal apakah seseorang kemungkinan terindikasi menderita penyakit diabetes.

METODE PENELITIAN

Metode penelitian yang digunakan pada penelitian ini adalah penelitian eksperimen evaluasi untuk mendapatkan algoritma terbaik dan paling akurat dalam memprediksi awal kemungkinan seseorang terindikasi penyakit diabetes. Adapun data yang digunakan pada penelitian ini adalah data diabetes yang didapatkan dari data publik *UCI Machine Learning Repository*. *UCI Machine Learning Repository* merupakan kumpulan *database*, *domain teory*, dan *generator data* yang digunakan oleh komunitas *Machine Learning* untuk analisis empiris algoritma *Data Mining*.

Dalam mendesain model penelitian eksperimen ini peneliti menerapkan permodelan standar yang digunakan pada *Data Mining* yaitu *Cross Industry Standard Process for Data Mining* (CRISP-DM) (David Olson, 2007). Adapun tahapannya yaitu: (1) *Business Understanding* (Pemahaman Bisnis) meliputi penetapan tujuan bisnis, penilaian situasi terkini, penetapan tujuan penggalian data dan pengembangan rencana proyek; (2) *Data Understanding* (Pemahaman Data) meliputi pengumpulan data awal, deskripsi data, eksplorasi data dan verifikasi kualitas data; (3) *Data Preparation* (Persiapan Data) meliputi proses *preprocessing* yaitu penyeleksian, pembersihan dan tranformasi data agar data dapat diproses ketahap pembuatan model; (4) *Modelling* (Pembuatan Model) meliputi penggambaran data dan penetapan hubungan serta analisis pengelompokan (mengidentifikasikan variabel mana yang berhubungan satu sama lain) dari algoritma *Decision Tree C4.5, Naive Bayes* dan *K-Nearest Neighbors* sehingga model-model lebih terinci yang sesuai dengan jenis data tersebut dapat diterapkan; (5) *Evaluation* (Evaluasi) meliputi proses evaluasi model menggunakan metode *Cross Validation* dan pengukuran *performance* menggunakan *Confusion Matrix, Compare ROC* dan *Paired T-Test*; (6) *Deployment* (Pelaksanaan) hasil data mining dapat digunakan baik untuk membuktikan hipotesis sebelumnya, ataupun untuk penemuan pengetahuan (pengidentifikasian hubungan yang tidak terduga dan bermanfaat). Melalui pengetahuan yang ditemukan dalam tahap awal proses CRISP-DM maka metode algoritma yang terbaik akan ditemukan yang dapat diterapkan pada kegiatan berbagai kebutuhan, termasuk memprediksi atau mengidentifikasi situasi-situasi penting.



Gambar 1. Kerangka pemikiran

HASIL DAN PEMBAHASAN

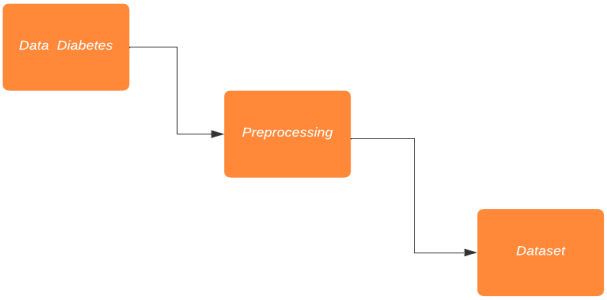
Dalam penelitian ini, dilakukan pengumpulan data diabetes dari data publik *UCI Machine Learning Repository*. Karakteristik data terdiri dari 17 variabel yaitu 16 sebagai variabel dependen (atribut x) dan 1 sebagai variabel independen (kelas y) dengan jumlah data sebanyak 520 data. Deskripsi dataset ditunjukkan pada Tabel 1.

Tabel 1. Deskripsi dataset diabetes

Variabel	Nama	Keterangan
Y	Diabetes	1. Positif 0. Negatif
X1	Umur Pasien	00 – 65 Tahun
X0	Jenis Kelamin	1. Laki - Laki 0. Perempuan
X3	*Polyuria* (Apakah pasien mengalami buang air kecil yang berlebihan)	1. Ya 0. Tidak
X4	*Polydipsia* (Apakah pasien mengalami rasa haus / minum berlebihan)	1. Ya 0. Tidak
X5	*Sudden Weight Loss* (Apakah pasien mengalami episode penurunan berat badan mendadak)	1. Ya 0. Tidak
X6	*Weakness* (Apakah pasien memiliki episode merasa lemah)	1. Ya 0. Tidak
X7	Polyphagia (Apakah pasien memiliki episode kelaparan yang berlebihan / ekstrim)	1. Ya 0. Tidak
X8	*Genital Thrush* (Apakah pasien memiliki infeksi ragi)	1. Ya 0. Tidak
X9	*Visual Blurring* (Apakah pasien memiliki episode penglihatan kabur)	1. Ya 0. Tidak
X10	*Itching* (Apakah pasien memiliki episode gatal)	1. Ya 0. Tidak
X11	*Irritability* (Apakah pasien memiliki episode lekas marah)	1. Ya

X10	*Delayed Healing* (Apakah pasien memiliki pemberitahuan penyembuhan tertunda ketika terluka)	0. Tidak 1. Ya
X13	*Partial Paresis* (Apakah pasien memiliki episode melemahnya otot / kelompok otot)	0. Tidak 1. Ya
X14	*Muscle Stiness* (Apakah pasien memiliki episode kekakuan otot)	0. Tidak 1. Ya
X15	*Alopecia* (Apakah pasien mengalami kerontokan rambut)	0. Tidak 1. Ya
X16	*Obesity* (Apakah pasien dapat dianggap obesitas)	0. Tidak 1. Ya

Data yang telah dikumpulkan dianalisa dan dilakukan pengolahan data awal sehingga atribut-atibut yang tidak relevan dibersihkan melalui tahap *preprocessing* (Gambar 2).

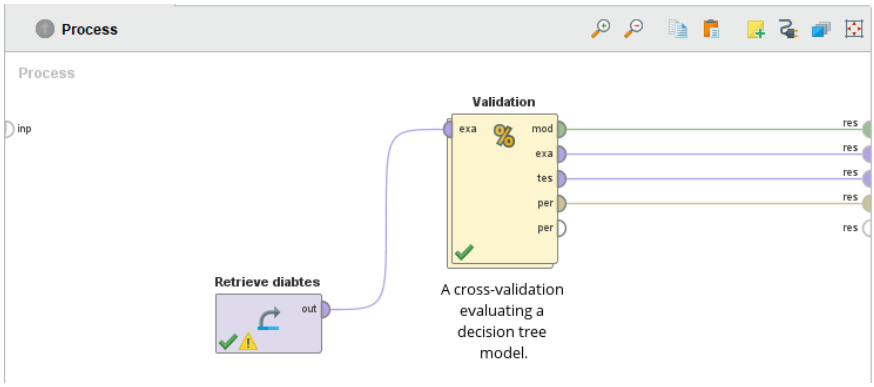


Gambar 2. Proses pengolahan data awal

Selanjutnya melakukan eksperimen permodelan data dan pengujian *performance* menggunakan *Cross Validation* dari algoritma *Decision Tree C4.5, Naive Bayes* dan *K-Nearest Neighbors*. Penggunaan *Cross Validation* akan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian sehingga didapatkan pilihan terbaik untuk mendapatkan hasil validasi yang akurat (Witten & Frank, 2018). Kemudian *performance* model dianalisa berdasarkan hasil *Confusion Matrix* yang berisi informasi tentang aktual (*actual*) dan prediksi (*predicted*) pada sistem klasifikasi dan dievaluasi dengan menggunakan data pada matriks (Febrian, 2011).

1. Pengujian menggunakan algoritma *Decision Tree C4.5*

Decision tree C4.5 digunakan untuk pengenalan pola termasuk juga pada pola-pola statistik dengan konsep dasarnya mengubah data menjadi pohon keputusan dengan aturan-aturannya (Permana & Patwari, 2021).



Gambar 3. Desain pengujian *Cross Validation Decesion Tree* C4.5

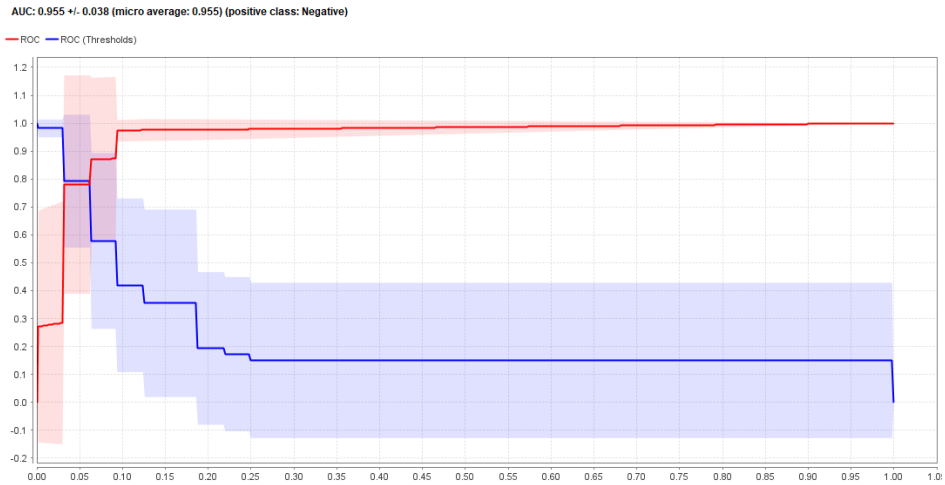
Adapun hasil dari pengujian menggunakan *Cross Validation* dari algoritma *Decision Tree C4.5* menunjukkan akurasi sebesar 96,35%. Berdasarkan *Confusion matrix* didapatkan bahwa dari 520 data, hasil prediksi benar positif diabetes dan kenyataannya benar positif sebanyak 308 data, hasil prediksi positif diabetes dan kenyataannya ternyata negatif sebanyak 7 data. Sedangkan hasil prediksi benar negatif diabetes dan kenyataannya benar negatif sebanyak 193 data, hasil prediksi negatif diabetes dan kenyataannya ternyata positif sebanyak 12 data.

accuracy: 96.35% +/- 2.47% (micro average: 96.35%)

	true Positive	true Negative	class precision
pred. Positive	308	7	97.78%
pred. Negative	12	193	94.15%
class recall	96.25%	96.50%	

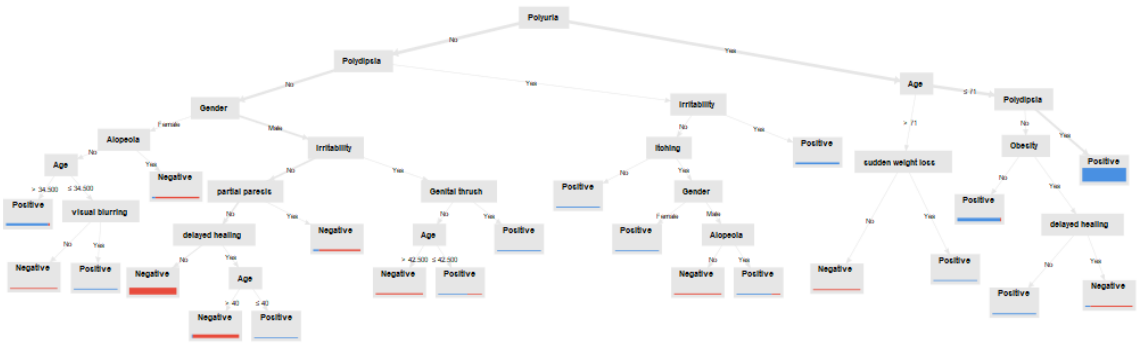
Gambar 4. Hasil performance akurasi *Decesion Tree* C4.5

Kemudian dari hasil pengujian juga didapatkan nilai AUC sebesar 0,955 sehingga membentuk grafik *area under curve* / *ROC Curve* seperti Gambar 5.



Gambar 5. Kurva ROC *decesion tree* C4.5

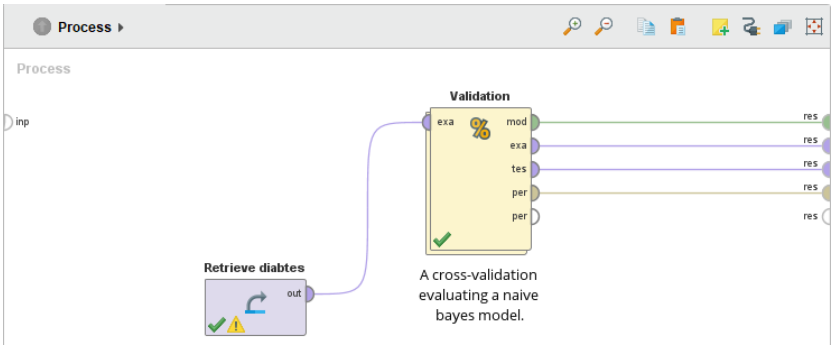
Metode algoritma *Decision Tree C4.5* menghasilkan model pohon keputusan yang didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada dan didapatkan atribut *polyuria* sebagai nilai *gain* tertinggi yaitu 0,263094024 sehingga dijadikan *node* akar.



Gambar 6. Model pohon keputusan *Decesion Tree* C4.5

2. Pengujian menggunakan algoritma *Naive Bayes*

Naive Bayes merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Leidiyana, 2011).



Gambar 7. Pengujian *Cross Validation* dari *Naive Bayes*

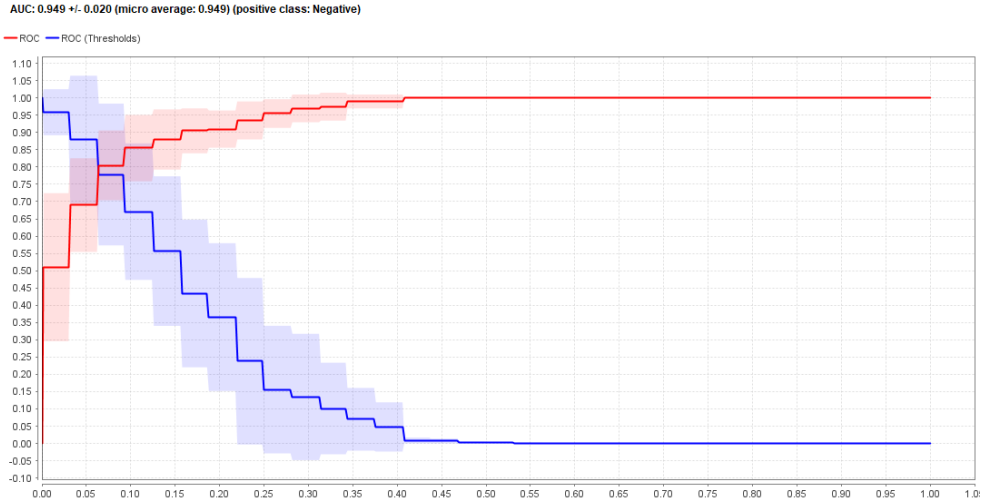
Hasil dari pengujian menggunakan *Cross Validation* dari algoritma *Naive Bayes* menunjukkan akurasi sebesar 87,69%. Berdasarkan *Confusion matrix* didapatkan bahwa dari 520 data, hasil prediksi benar positif diabetes dan kenyataannya benar positif sebanyak 276 data, hasil prediksi positif diabetes dan kenyataannya ternyata negatif sebanyak 20 data. Sedangkan hasil prediksi benar negatif diabetes dan kenyataannya benar negatif sebanyak 180 data, hasil prediksi negatif diabetes dan kenyataannya ternyata positif sebanyak 44 data.

accuracy: 87.69% +/- 4.98% (micro average: 87.69%)

	true Positive	true Negative	class precision
pred. Positive	276	20	93.24%
pred. Negative	44	180	80.36%
class recall	86.25%	90.00%	

Gambar 8. Hasil performance akurasi *Naive Bayes*

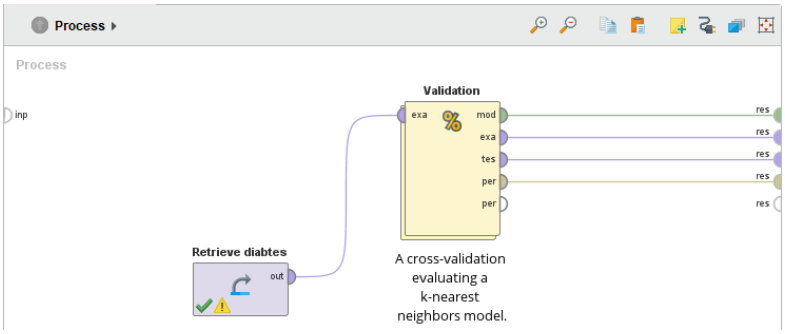
Kemudian dari hasil pengujian juga didapatkan nilai AUC sebesar 0,949 sehingga membentuk grafik *area under curve* / *ROC Curve* sebagaimana ditunjukkan pada Gambar 9.



Gambar 9. Kurva ROC *Naive Bayes*

3. Pengujian menggunakan algoritma *K-Nearest Neighbors*

K-Nearest Neighbors merupakan suatu metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.



Gambar 10. Pengujian *Cross Validation* dari *K-Nearest Neighbors*

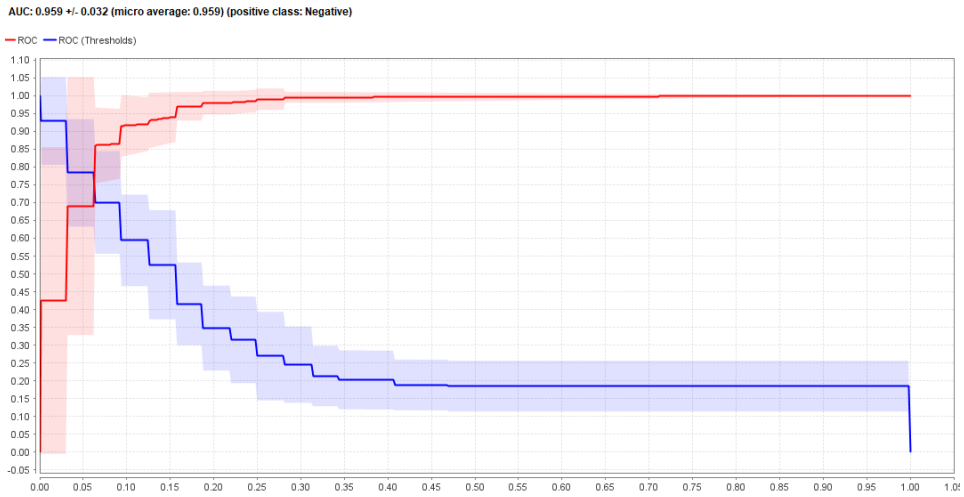
Hasil dari pengujian menggunakan *Cross Validation* dari algoritma *K-Nearest Neighbors* menunjukkan akurasi sebesar 89,62%. Berdasarkan *Confusion matrix* didapatkan bahwa dari 520 data, hasil prediksi benar positif diabetes dan kenyataannya benar positif sebanyak 281 data, hasil prediksi positif diabetes dan kenyataannya ternyata negatif sebanyak 15 data. Sedangkan hasil prediksi benar negatif diabetes dan kenyataannya benar negatif sebanyak 185 data, hasil prediksi negatif diabetes dan kenyataannya ternyata positif sebanyak 39 data.

accuracy: 89.62% +/- 3.87% (micro average: 89.62%)

	true Positive	true Negative	class precision
pred. Positive	281	15	94.93%
pred. Negative	39	185	82.59%
class recall	87.81%	92.50%	

Gambar 11. Hasil *performance* akurasi *K-Nearest Neighbors*

Kemudian dari hasil pengujian juga didapatkan nilai AUC sebesar 0,959 sehingga membentuk grafik *area under curve* / *ROC Curve* sebagaimana Gambar 12.



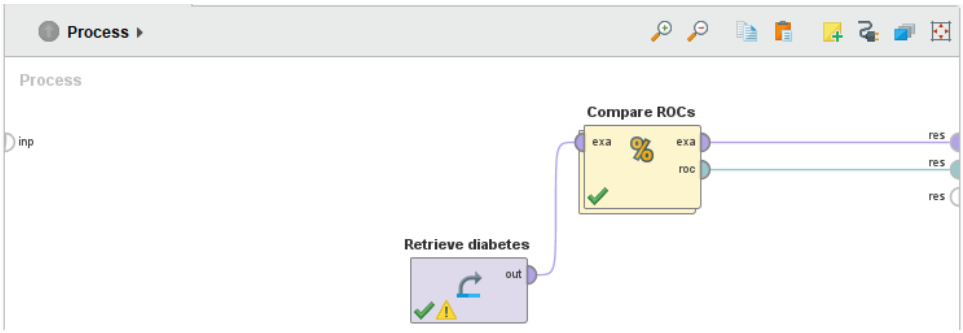
Gambar 12. Kurva ROC *K-Nearest Neighbors*

Hasil analisis perbandingan *performance Confusion Matrix* dari algoritma *Decision Tree C4.5, Naive Bayes* dan *K-Nearest Neighbors* maka dapat dirangkum ke dalam Tabel 2.

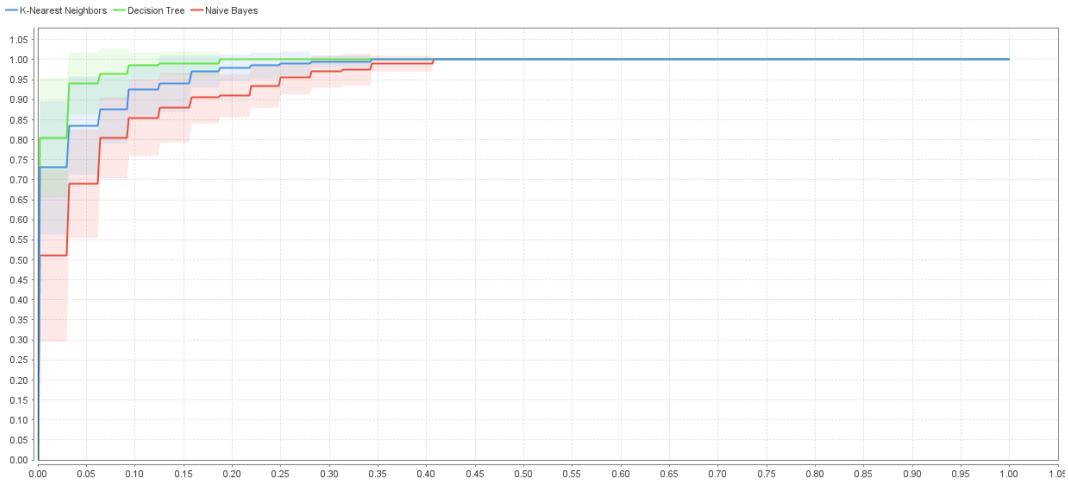
Tabel 2. Komparasi *performance Confusion Matrix*

Algoritma	Akurasi	Nilai AUC
Decision Tree C4.5	96,35%	0,955
Naive Bayes	87,69%.	0,949
K-Nearest Neighbors	89,62%	0,959

Berdasarkan perbandingan *performance* pada Tabel 2 menunjukkan bahwa algoritma *Decision Tree C4.5* memiliki tingkat akurasi prediksi tertinggi diikuti algoritma *K-Nearest Neighbors* lalu *Naive Bayes*. Selain pengujian *performance* dengan menggunakan *Confusion Matrix*, pengujian juga dilakukan menggunakan *Compare ROCs* agar diketahui perbandingan tingkat kehandalan dari algoritma *Decision Tree C4.5, Naive Bayes* dan *K-Nearest Neighbors.*



Gambar 13. Desain model komparasi menggunakan *Compare ROCs*

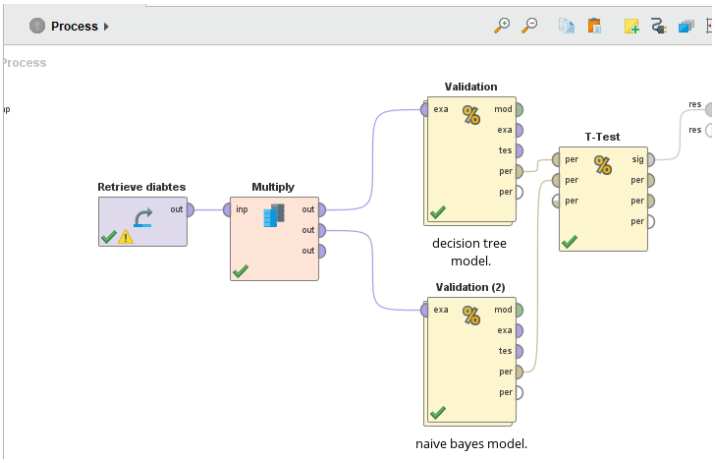


Gambar 14. Grafik komparasi kurva ROC

Menurut Gorunescu (2015), *performance* kehandalan nilai AUC dapat diklasifikasikan menjadi lima kelompok yaitu: (1) 0.90 – 1.00 = *Excellent Classification*; (2) 0.80 – 0.90 = *Good Classification*; (3) 0.70 – 0.80 = *Fair Classification*; (4) 0.60 – 0.70 = *Poor Classification*; dan (5) 0.50 – 0.60 = *Failure*.

Berdasarkan hasil perbandingan dari *Compare ROCs*, maka baik *Decision Tree C4.5, Naive Bayes* maupun *K-Nearest Neighbors* termasuk dalam kelompok *excellent*

classification. Selain pengujian dengan *Confusion Matrix* dan *Compare ROCs*, pengujian model juga menggunakan pengujian *Paired T-Test*. Pengujian *T-Test* merupakan metode pengujian berdasarkan hipotesis menggunakan satu objek penelitian dengan dua perlakuan yang berbeda (Hastuti, 1994).



Gambar 15. Desain pengujian *performance* T-Test

Pada pengujian statistik menggunakan *T-Test* ini, Dalam pengujian ini, akan dibandingkan dua metode algoritma secara bergantian sehingga didapatkan hasil komparasi pada Tabel 3.

Tabel 3. Hasil komparasi uji statistik T-Test

	Decision Tree C4.5	*Naive Bayes*	*K-Nearest Neighbors*
Decision Tree C4.5		**0,000**	**0,000**
Naive Bayes	**0,000**		0,439
K-Nearest Neighbors	**0,000**	0,293	

Hasil pengujian statistik *Paired T-Test* pada Tabel 3 dapat dianalisis bahwa algoritma *Decision Tree C4.5* sangat dominan atau signifikan karena memiliki nilai probabilitas < 0.05 terhadap algoritma lainnya.

Tabel 4. Hasil seluruh metode pengujian

Metode	***Decision Tree C4.5***	***Naive Bayes***	***K-Nearest Neighbors***
Confusion Matrix (akurasi)	96,35%	87,69%.	89,62%
Compare ROCs (nilai AUC)	0,955	0,949	0,959
Paired T-Test (nilai uji statistic)	Dominan	Tidak Dominan	Tidak Dominan

Berdasarkan hasil seluruh metode pengujian yang digunakan pada Tabel 4, maka dapat disimpulkan bahwa algoritma yang paling akurat dalam memprediksi awal kemungkinan seseorang terindikasi penyakit diabetes adalah algoritma *Decision Tree C4.5* karena memiliki akurasi tertinggi sebesar 96,35% dan sangat dominan terhadap algoritma lainya berdasarkan hasil uji statistik serta termasuk dalam kategori *excellent classification*.

KESIMPULAN

Berdasarkan komparasi hasil pengujian *Confusion Matrix, Compare ROCs*, dan *Paired T-Test* dari algoritma *Decision Tree C4.5*, *Naive Bayes* dan juga *K-Nearest Neighbors* dalam memprediksi awal kemungkinan seseorang terindikasi penyakit diabetes, maka dapat disimpulkan bahwa algoritma *Decision Tree C4.5* merupakan algoritma yang paling akurat dan paling dominan terhadap algoritma lainya serta termasuk ke dalam kategori *excellent classification* sehingga dengan demikian algoritma *Decision Tree C4.5* dapat memberikan pemecahan untuk permasalahan penentuan apakah seseorang terindikasi menderita penyakit diabetes terutama pada fase asimtomatik sebagai pendukung pengambilan keputusan sebelum pemeriksaan lebih lanjut.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada tim peneliti Data Mining Fakultas Teknologi Informasi, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin. Penelitian ini didukung dan disponsori oleh Yayasan Universitas Islam Kalimantan Muhammad Arsyad Al Banjari Banjarmasin Nomor: 45/UNISKA-PUSLIT/II/2022.

DAFTAR PUSTAKA

Anggraeni, D., & Ramadhani. (2018). Analisa decision tree untuk prediksi diagnosa diabetes mellitus. *Seminar Nasional Royal (Senar)*, 1(1), 153-158.

Candra Permana, B. A., & Dewi Patwari, I. K. (2021). Komparasi metode klasifikasi data mining decision tree dan naïve bayes untuk prediksi penyakit diabetes. *Infotek : Jurnal Informatika Dan Teknologi*, 4(1), 63–69. https://doi.org/10.29408/jit.v4i1.2994.

Detikhealth. (2019). *Mengenal Penyakit Diabetes, Penyebab dan Cara Mengatasinya*. https://health.detik.com/berita-detikhealth/d-4468242/mengenal-penyakit-diabetes-penyebab-dan-cara-mengatasinya.

Edy. (2012). Penerapan Algoritma C4.5 Dengan Seleksi Atribut Berbasis Algoritma Genetika Dalam Diagnosa Penyakit Jantung. [Tesis]. Jakarta: Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.

Eldridge, B. L. (2022). *The Significance of Diagnosing a Disease as Asymptomatic*. Verywell Health. https://www.verywellhealth.com/asymptomatic-definition-importance-and-controversy-2249055.

Febrian, F. (2011). Algoritma Klasifikasi Data Mining Pada Akseptasi Data Fakultatif. [Skripsi]. Jakarta: Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.

Gorunescu, F. (2015). Data Mining: Concepts, Models and Techniques. In *Mining of Massive Datasets*. Heidelberg: Springer Berlin.

Handayanna, F. (2012). Penerapan *Particle Swarm Optimization* Untuk Seleksi Atribut Pada Metode *Support Vector*. [Tesis]. Jakarta: Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri.

Hastuti, K. (1994). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. *Seminars in Neurology*, 14(1), 241–249. https://doi.org/10.2307/j.ctv11hppt6.3.

Witten, I. H., & Eibe Frank, M. A. H. (2018). Data mining: Practical Machine Learning Tools and Techniques. *Angewandte Chemie International Edition*, 6(11), 951–952.

International Diabetes Federation. (2021). *Diabetes Facts and Figures*. https://www.idf.org/aboutdiabetes/what-is-diabetes/glossary.html.

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *Advances in Intelligent Systems and Computing*, 992, 113–125. https://doi.org/10.1007/978-981-13-8798-2_12.

Kementerian Kesehatan RI. (2021). *Penyakit Diabetes Melitus*. http://p2ptm.kemkes.go.id/infographic-p2ptm/penyakit-diabetes-melitus.

Lasut, D. (2012). Prediksi Loyalitas Pelanggan Pada Perusahaan Penyedia Layanan Multimedia Dengan Algoritma C4.5 berbasis *Particle Swarm Optimization*. [Skripsi]. Jakarta: Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.

Leidiyana, H. (2011). Komparasi Algoritma Klasifikasi Data Mining Dalam Penentuan Resiko Kredit

Kepemilikan Kendaraan Bemotor. [Skripsi]. Jakarta: Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri.

Mustafa, M. S., & Simpen, I. W. (2019). Implementasi algoritma k-nearest neighbor (knn) untuk memprediksi pasien terkena penyakit diabetes pada Puskesmas Manyampa Kabupaten Bulukumba. *Seminar Ilmiah Sistem Informasi dan Teknologi Informasi*, VIII(1), 1–10.

Olson, D. L., & Shi, Y. (2007). *Introduction to Business Data Mining*. United States: McGraw-Hill.

Rumini, & Nasruddin, A. (2021). Prediksi awal penyakit diabetes mellitus menggunakan algoritma naive bayes. *Jurnal ICT: Information Communication & Technology,* 20(2), 246–253.

Suradiradja, K. H. (2018). Deteksi Transaksi Pencucian Uang Dengan Algoritma Klasifikasi C4.5. *Jurnal Teknologi Informasi ESIT*, XII(01), 62-66.