BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

Komparasi Algoritma Data Mining untuk Klasifikasi Penyakit Diabetes

Ivandari[1], Much. Rifqi Maulana[2], Muhammad Faizal Kurniawan[2], M Adib Al Karomi[2,*]

[1]Program Studi Sistem Informasi, STMIK Widya Pratama, Pekalongan, Indonesia
[2]Program Studi Teknik Informatika, STMIK Widya Pratama, Pekalongan, Indonesia
Email: [1]ivandarialkaromi@gmail.com, [2]rifqi@stmik-wp.ac.id, [3]mfaizalkurniawan@gmail.com, [3,*]adib@alkaromi.id
Email Penulis Korespondensi: adib@alkaromi.id

Abstrak−Diabetes merupakan salah satu penyakit tidak menular mematikan yang dapat menyerang manusia. Dari data World Health Organization (WHO) diabetes telah membunuh setidaknya 2 juta manusia sepanjang tahun 2019. Pencatatan dari tiap fase dan kondisi pasien diabetes banyak dilakukan guna menunjang penelitian. Salah satu catatan terupdate dari pasien diabetes adalah early stage diabetes risk prediction dataset. Dataset ini dirilis uci repository pada akhir 2020 oleh Rumah Sakit Diabetes di Bangladesh. Klasifikasi dalam data mining merupakan ilmu yang dapat mengekstraksi data untuk mencari pola atau model data guna mendapatkan pengetahuan baru. Beberapa algoritma klasifikasi yang banyak digunakan dan terbukti dapat menangani data yang besar antara lain K-NN, Naïve Bayes, serta Decission Tree. Penelitian ini melakukan komparasi ketiga algoritma tersebut untuk klasifikasi early stage diabetes risk prediction dataset. Dari hasil penelitian decision tree merupakan algoritma terbaik untuk klasifikasi dataset diabetes dengan tingkat akurasi sebesar 95,96%. Berikutnya algoritma KNN dengan tingkat akurasi 92,5%. Sedangkan naïve bayes tercatat hanya menghasilkan tingkat akurasi sebesar 86,92%. Dari komparasi ini diketahui bahwa decision tree merupakan algoritma terbaik untuk klasifikasi early stage diabetes risk prediction dataset dengan tingkat akurasi 95,96%.

Kata Kunci: Decission Tree; Naïve Bayes; K-NN

Abstract−Diabetes is one of the deadly non-communicable diseases that can attack humans. According to data from the World Health Organization (WHO), diabetes has killed at least 2 million people throughout 2019. Many recordings of each phase and condition of diabetes patients are done to support research. One of the most updated records of diabetes patients is the early stage diabetes risk prediction dataset. This dataset was released by the uci repository in late 2020 by the Diabetes Hospital in Bangladesh. Classification in data mining is a science that can extract data to look for patterns or data models to gain new knowledge. Several classification algorithms that are widely used and proven to be able to handle large data include K-NN, Naïve Bayes, and Decission Tree. This study compares the three algorithms to classify early stage diabetes risk prediction dataset. From the research results, the decision tree is the best algorithm for classifying diabetes datasets with an accuracy rate of 95.96%. Next is the KNN algorithm with an accuracy rate of 92.5%. Meanwhile, naïve Bayes only produces an accuracy rate of 86.92%. From this comparison it is known that the decision tree is the best algorithm for classifying the early stage diabetes risk prediction dataset with an accuracy rate of 95.96%.

Keywords: Decission Tree; Naïve Bayes; K-NN

1. PENDAHULUAN

Diabetes merupakan penyakit dengan resiko kematian yang tinggi. Dalam tahun 2019 WHO mencatat setidaknya 2 juta kematian akibat diabetes [1]. Faktor utama pemicu diabetes adalah banyaknya kadar glukosa dalam darah sehingga tubuh tidak dapat mengontrol kadar glukosa dalam darah [2]. Diabetes yang berbahaya bukan merupakan penyakit genetic yang dapat diturunkan orang tua ke anak keturunannya. Factor utama munculnya diabetes adalah pola makan yang tidak sehat. Sampai saat ini pengobatan total untuk menyembuhkan diabetes belum ditemukan. Pasien hanya diberikan insulin karena tubuh tidak dapat memproduksi insulin dengan baik. Penanganan dini terhadap penyakit diabetes dapat mengurangi resiko yang lebih parah terhadap pasien.

Pesatnya perkembangan teknologi dan komputer memiliki peranan penting dalam pencatatan dan analisa data di semua bidang. Data mining merupakan bidang ilmu yang dapat menemukan pola dari sebuah data [3]. Konsep dari data mining adalah menemukan pengetahuan baru dari hasil perhingan statistic data yang sudah ada [4]. Perhitungan dalam proses data mining akan menghasilkan tingkat akurasi dan dapat dibandingkan dengan menggunakan banyak model algoritmik [5]. Beberapa model perhingan algoritma data mining dapat menangani jenis dan model data tertentu. Salah satu fungsi utama data mining adalah klasifikasi. Dalam klasifikasi data awal atau data training sangat mempengaruhi hasil klasifikasi. Dari data awal tersebut nantinya akan muncul pengetahuan baru atau pola data yang dapat dijadikan acuan untuk menentukan kebijakan selanjutnya. Beberapa algoritma klasifikasi yang terbukti baik dan dapat menangani data yang besar antara lain naïve bayes, k-nn, serta decision tree [6].

Klasifikasi data mining terbukti dapat digunakan dalam semua bidang, salah satunya bidang kesehatan. Proses klasifikasi ini dapat mendapatkan pola baru dari sebuah dataset. Hasil klasifikasi dapat digunakan antara lain untuk deteksi dini penyakit ginjal [7], deteksi penyakit kanker payudara [8] deteksi dini diabetes [9] serta beberapa penyakit lain. Klasifikasi dataset di bidang kesehatan nantinya dapat digunakan sebagai pendukung keputusan.

Pada tahun 2021 Saloni Kumari dari Department of Electronics and Communication Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India melakukan komparasi algoritma untuk diabetes mellitus. Dalam penelitian ini digunakan berbagai macam algoritma klasifikasi untuk dibandingkan. Beberapa algoritma klasifikasi yang digunakan antara lain logistic regression, KNN, SVM, naïve bayes, soft voting classifier dan beberapa algoritma lain. Dalam penelitian ini soft voting classifier menghasilkan tingkat akurasi terbaik yaitu 79,08% [10].

Pada awal 2023 Cesar Carpinteiro dari LASI Research Center, University of Minho, Portugal juga melakukan studi komparasi diabetes serupa. Beberapa algoritma yang digunakan dalam klasifikasi ini antara lain logistic regression, KNN, SVM, MLP, GBM dan beberapa algoritma klasifikasi lainnya. Dalam penelitian carpenteiro ini didapatkan hasil 3

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

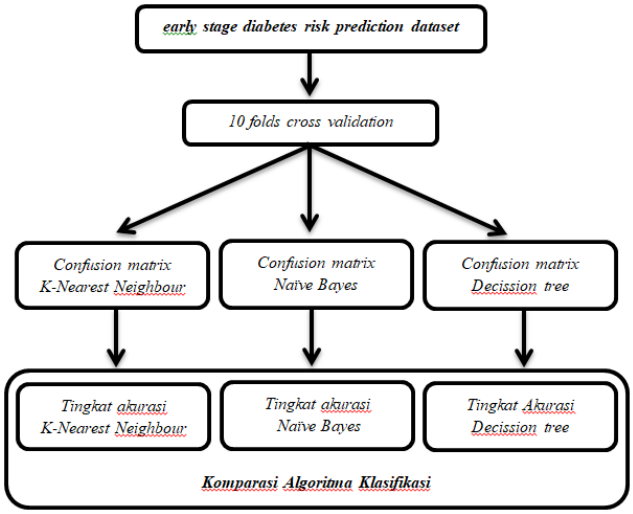
algoritma dengan tingkat akurasi tertinggi. Algoritma tersebut adalah SVM dengan tingkat akurasi 80,00%, dilanjutkan oleh MLP dengan tingkat akurasi 80,2%. Algoritma terbaik dalam proses klasifikasi ini adalah Gradient Boost Machine (GBM) dengan tingkat akurasi tertinggi yaitu 80,4% [11].

Dalam penelitian ini menggunakan dataset diabetes terupdate dari uci repository. Uci repository merupakan portal pengelola dataset terkemuka yang banyak digunakan peneliti untuk melakukan pengujian algoritma. Dataset dapat diakses di https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset. Dalam dataset ini terdapat 17 attribute dan 520 record. Perhitungan dilakukan dengan menggunakan aplikasi rapid miner. Dalam proses perhitungan algoritma dilakukan validasi menggunakan 10 folds cross validation. Proses ini memungkinkan dataset dibagi menjadi 10 bagian, 9 bagian diantaranya digunakan sebagai data training dan 1 bagian digunakan sebagai data testing. Proses ini dilakukan berulang sampai semua bagian mendapatkan porsi menjadi data testing. Proses terakhir adalah pengujian dengan menggunakan confusion matrix. Matrix ini menghasilkan tingkat akurasi klasifikasi semua algoritma. Tingkat akurasi didapatkan dari prosentase hasil perhitungan data yang sesuai dengan kondisi sesungguhnya. Semakin banyak nilai tingkat akurasi artinya algoritma atau metode tersebut semakin baik.

Dengan adanya model algoritma yang tepat untuk data diabetes diharapkan penelitian berikutnya dapat dibuat sebuah system pendukung keputusan deteksi dini penyakit diabetes. sistem tersebut dapat menganalisa lebih dini perilaku dan pola hidup manusia agar dapat menurunkan resiko keparahan akibat penyakit diabetes.

2. METODOLOGI PENELITIAN

Penelitian ini tergolong dalam penelian eksperimental yang dilakukan dengan beberapa tahapan. Tahapan pertama adalah pengumpulan data, dilanjutkan dengan analisisi data. Kemudian proses penelitian dilakukan dengan menggunakan aplikasi bantu rapid miner. Dalam tahapan ini dilakukan validasi menggunakan 10 folds cross validation serta proses evaluasi dilakukan dengan menggunakan convussion matrix. Gambar 1 merupakan tahapan penelitian yang dilakukan. Secara lebih terperinci proses tahapan penelitian dilakukan sebagaimana berikut:



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Dalam penelitian ini digunakan data dari uci repository. Uci repository merupakan portal penyedia data yang banyak digunakan peneliti untuk pengujian algoritma. Didalamnya terdapat lebih dari 600 dataset yang telah digunakan oleh jutaan peneliti di dunia. Dataset prediksi resiko diabetes ini didonasikan pada 7 November 2020 [12]. Data ini dikumpulkan atas persetujuan dokter dengan melakukan kuisioner kepada 520 pasien Rumah Sakit Diabetes Sylhet, Bangladesh. Dataset ini sudah tertata rapi dalam file csv dan tergolong open access serta dapat diunduh di https://archive.ics.uci.edu/static/public/529/early+stage+diabetes+risk+prediction+dataset.zip. Tabel 1 merupakan metadata dari early stage diabetes risk prediction dataset.

Tabel 1. Metadata early stage diabetes risk prediction dataset

Role	Name	Type	Statistics	Range	Missings
label	class	binominal	mode = Positive (320), least = Negative (200)	Positive (320), Negative (200)	0.0
regular	Age	integer	avg = 48.029 +/- 12.151	[16.000 ; 90.000]	0.0
regular	Gender	binominal	mode = Male (328), least = Female (192)	Male (328), Female (192)	0.0
regular	Polyuria	binominal	mode = No (262), least = Yes (258)	No (262), Yes (258)	0.0
regular	Polydipsia	binominal	mode = No (287), least = Yes (233)	Yes (233), No (287)	0.0

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

Role	Name	Type	Statistics	Range	Missings
regular	sudden weight loss	binominal	mode = No (303), least = Yes (217)	No (303), Yes (217)	0.0
regular	weakness	binominal	mode = Yes (305), least = No (215)	Yes (305), No (215)	0.0
regular	Polyphagia	binominal	mode = No (283), least = Yes (237)	No (283), Yes (237)	0.0
regular	Genital thrush	binominal	mode = No (404), least = Yes (116)	No (404), Yes (116)	0.0
regular	visual blurring	binominal	mode = No (287), least = Yes (233)	No (287), Yes (233)	0.0
regular	Itching	binominal	mode = No (267), least = Yes (253)	Yes (253), No (267)	0.0
regular	Irritability	binominal	mode = No (394), least = Yes (126)	No (394), Yes (126)	0.0
regular	delayed healing	binominal	mode = No (281), least = Yes (239)	Yes (239), No (281)	0.0
regular	partial paresis	binominal	mode = No (296), least = Yes (224)	No (296), Yes (224)	0.0
regular	muscle stiffness	binominal	mode = No (325), least = Yes (195)	Yes (195), No (325)	0.0
regular	Alopecia	binominal	mode = No (341), least = Yes (179)	Yes (179), No (341)	0.0
regular	Obesity	binominal	mode = No (432), least = Yes (88)	Yes (88), No (432)	0.0

Dataset ini memiliki 17 atribut data. Salah satu atribut yaitu class merupakan atribut label. Atribut label ini merupakan atribut yang nantinya akan dicari untuk diprediksi untuk pasien atau record baru. Selain atribut tersebut kesemuanya merupakan atribut regular yang digunakan untuk proses perhitungan. Jenis kesemua atribut sama yaitu binominal atau dua pilihan yaitu yes dan no. kecuali atribut usia (age) yang berjenis integer atau angka yang memiliki nilai. Tabel 2 merupakan dataset yang digunakan.

Tabel 2. Early stage diabetes risk prediction dataset

No	Class	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity
1	Positive	40.0	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes
2	Positive	58.0	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No
3	Positive	41.0	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No
4	Positive	45.0	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No
5	Positive	60.0	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
6	Positive	55.0	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes
7	Positive	57.0	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No
8	Positive	66.0	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No	No
9	Positive	67.0	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
10	Positive	70.0	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No
11	Positive	44.0	Male	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	No	Yes	Yes	No
12	Positive	38.0	Male	Yes	Yes	No	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No
13	Positive	35.0	Male	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No
14	Positive	61.0	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes
15	Positive	60.0	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No
16	Positive	58.0	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
508	Negative	60.0	Male	No	No	Yes	No	No	No	No	No	No	No	No	No	No	Yes
509	Negative	58.0	Male	No	No	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	No
510	Negative	54.0	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

No	Class	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity
511	Negative	67.0	Male	No	No	No	Yes	No	No	No	Yes	No	Yes	No	No	Yes	No
512	Negative	66.0	Male	No	No	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No
513	Negative	43.0	Male	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No
514	Positive	62.0	Female	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	No	No	Yes
515	Positive	54.0	Female	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	No	No	No
516	Positive	39.0	Female	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	No	No	No
517	Positive	48.0	Female	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No
518	Positive	58.0	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	Yes
519	Negative	32.0	Female	No	No	No	Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	No
520	Negative	42.0	Male	No	No	No	No	No	No	No	No	No	No	No	No	No	No
508	Negative	60.0	Male	No	No	Yes	No	No	No	No	No	No	No	No	No	No	Yes

2.2 Analisis Data

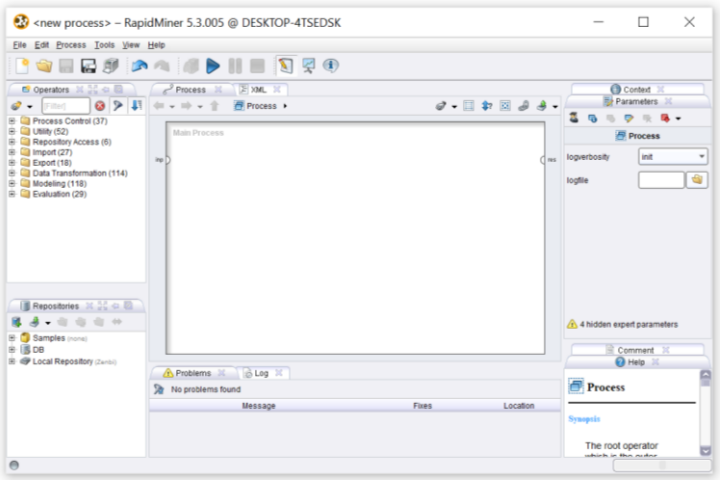
Dataset yang digunakan memiliki 17 atribut dengan 16 atribut regular dan 1 atribut label. Sebagian besar atribut memiliki tipe atribut binominal. Tipe ini hanya memiliki 2 kemungkinan didalamnya. Hanya satu atribut yang memiliki tipe atribut integer yaitu atribut age (usia). Jumlah record dalam dataset ini adalah 520 tanpa adanya missing value didalamnya. Pada metadata di tabel 1 terlihat bahwa data dalam kondisi seimbang. Artinya ketimpangan data tidak terjadi untuk semua atribut data yang ada. Selanjutnya dalam proses perhitungan semua atribut data akan digunakan utnuk proses menggunakan ketiga algoritma.

2.3 Proses Perhitungan Algoritma

Proses perhitungan untuk ketiga algoritma dilakukan dengan menggunakan aplikasi bantu rapid miner. Aplikasi ini dapat melakukan input data dari database excel, csv dan yang lainnya. Aplikasi rapid miner merupakan aplikasi open source yang dapat digunakan peneliti untuk melakukan proses perhitungan data mining. Gambar 2 merupakan tampilan rapid miner. Sedangkan gambar 3 adalah lembar kerja utama rapid miner.

Gambar 2. Tampilan utama rapid miner

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

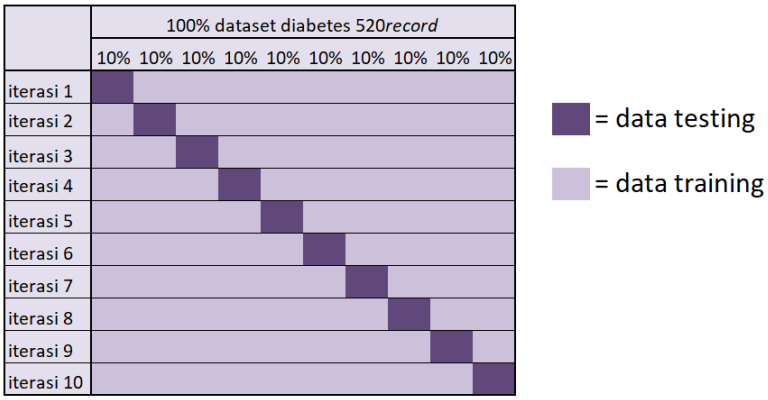


Gambar 3. Lembar kerja rapid miner

Dalam proses perhitungan dilakukan dengan 2 tahapan yaitu validasi dan evaluasi yang masing masing sebagaimana berikut:

2.3.1 Validasi

Validasi merupakan proses dimana semua data dipastikan digunakan sebagaimana mestinya dalam proses penelitian. Proses validasi yang umum digunakan dalam klasifikasi adalah cross validation [13], penelitian ini menggunakan 10 folds cross validation. Proses dalam 10 folds cross validation adalah dengan membagi sama rata dataset menjadi 10 bagian. Kemudian 1 diantaranya digunakan sebagai data testing, sedangkan 9 yang lainnya digunakan sebagai data training. Proses ini diulang sampai 10 kali, sampai semua bagian data tersebut berkesempatan 1 kali digunakan sebagai data testing[14]. Gambar 4 merupakan representasi dari 10 folds cross validation.



Gambar 4. Representasi 10 folds cross validation

2.3.2 Evaluasi

Evaluasi dilakukan untuk menilai dan membandingkan hasil perhitungan. Proses evaluasi yang banyak digunakan dalam penelitian klasifikasi adalah menggunakan confusion matrix atau matrix kebingungan. Matrix ini memperhitungkan jumlah semua record data testing dari hasil klasifikasi yang dilakukan. Perhitungannya adalah jumlah data record yang sesuai dengan label aslinya dibagi dengan keseluruhan jumlah perhitungan [15]. Dari proses perhitungan setiap iterasi memiliki nilai yang berbeda. Hasil prosentase matrix ini adalah rata rata dari 10 iterasi yang ada. Gambar 5 merupakan representasi proses perhitungan matrix kebingungan[16].

CLASSIFICATION		PREDICTED CLASS	
		Class = YES	**Class = NO**
OBSERVED CLASS	**Class = YES**	*a* (*true positive*-TP)	*b* (*false negative* -FN)
	Class = NO	*c* (*false positive*-FP)	*d* (*true negative*-TN)

Gambar 5. Representasi confusion matrix [16]

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

Nilai tingkat akurasi dari sebuah algoritma dapat dihitung dengan rumus (1) berikut:

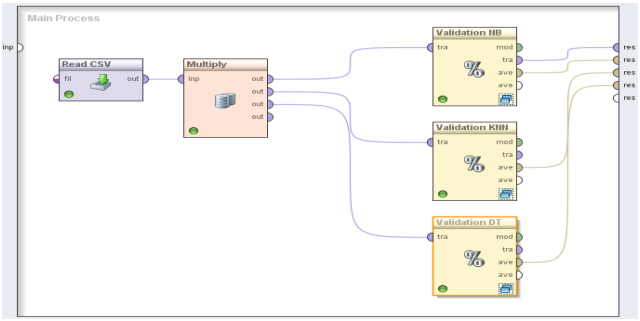
$$accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

a = (true positive-TP) = adalah jumlah prediksi YES dan hasil klasifikasi label sebenarnya YES = klasifikasi benar
b = (false negative-TP) = adalah jumlah prediksi NO dan hasil klasifikasi label benar YES = klasifikasi salah
c = (false positive-TP) = adalah jumlah prediksi YES dan hasil klasifikasi label benar NO = klasifikasi salah
d = (true negative-TP) = adalah jumlah prediksi NO dan hasil klasifikasi label benar NO = klasifikasi benar

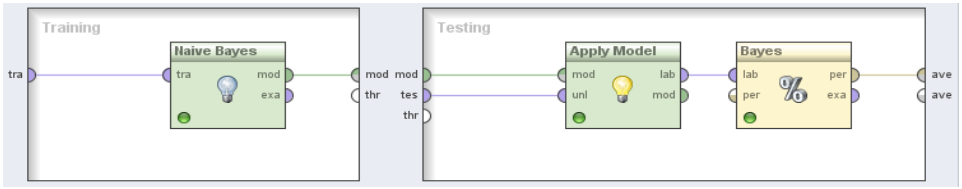
3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

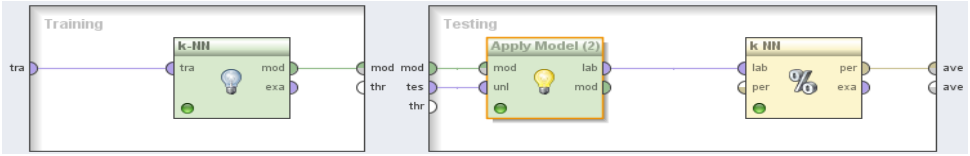
Penelitian menggunakan rapid miner dilakukan dengan drag and drop dataset, serta model validasi dan evaluasi serta algoritma yang digunakan. Gambar 6 merupakan proses perhitungan pada lembar kerja utama dalam penelitian ini. Didalamnya terdapat dataset dengan tipe csv. Dikarenakan komparasi dilakukan dengan menggunakan 3 jenis algoritma, maka dibutuhkan konektor multiply output agar proses validasi memiliki porsi data training dan data testing yang sama untuk ketiga algoritma. Dalam proses validasi terdapat perhitungan confusion matrix dengan menggunakan masing masing algoritma. Gambar 7 merupakan proses perhitungan naïve bayes, sedangkan gambar 8 merupakan proses perhitungan KNN, dilanjutkan gambar 9 merupakan proses perhitungan Decission tree.



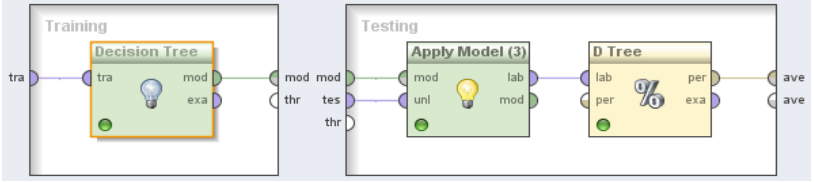
Gambar 6. Proses penelitian dengan rapid miner



Gambar 7. Perhitungan naïve bayes dengan rapid miner



Gambar 8. Perhitungan KNN dengan rapid miner



Gambar 9. Perhitungan decission tree dengan rapid miner

3.1.1 Tingkat akurasi naïve bayes

Naïve bayes merupakan model algoritma dengan dasar perhitungan probabilitas dan merupakan pengembangan dari terema bayes [17]. Dalam model teori ini atribut nominal dapat digunakan dan berjalan dengan baik. Sebaliknya, teori ini akan menjadi lemah apabila atribut yang digunakan memiliki tipe numeric atau integer. Dengan menggunakan naïve bayes klasifikasi diabetes dalam penelitian ini mendapatkan tingkat akurasi sebesar 86,92%. Gambar 10 merupakan tingkat akurasi confusion matrix dari naïve bayes yang dihasilkan dari penelitian ini

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

Multiclass Classification Performance Annotations

Table View Plot View

accuracy: 86.92% +/- 3.31% (mikro: 86.92%)

	true Positive	true Negative	class precision
pred. Positive	275	23	92.28%
pred. Negative	45	177	79.73%
class recall	85.94%	88.50%	

Gambar 10. Perhitungan naïve bayes dengan rapid miner

3.1.2 Tingkat akurasi K-Nearest Neighbour

KNN adalah algoritma klasik yang diperkenalkan pertama kali pada tahun 1967 [18]. Selanjutnya banyak pengembangan yang membuat algoritma KNN diminati dan banyak digunakan dalam proses klasifikasi. Secara sederhana KNN mencari kedekatan dengan cara mencocokkan kasus baru dengan kasus sebelumnya dengan pembobotan tertentu [19]. Keberhasilan algoritma KNN sangat dipengaruhi oleh banyaknya atribut yang relevan yang digunakan dalam proses klasifikasi [20]. Penelitian ini menggunakan keseluruhan dari 16 atribut regular yang ada dan menghasilkan tingkat akurasi sebesar 92,5% untuk algoritma KNN. Gambar 11 merupakan tingkat akurasi confusion matrix dari knn yang dihasilkan dari penelitian ini.

Multiclass Classification Performance Annotations

Table View Plot View

accuracy: 92.50% +/- 4.42% (mikro: 92.50%)

	true Positive	true Negative	class precision
pred. Positive	296	15	95.18%
pred. Negative	24	185	88.52%
class recall	92.50%	92.50%	

Gambar 11. Perhitungan knn dengan rapid miner

3.1.3 Tingkat akurasi decision tree

Decission tree merupakan model yang banyak digunakan dalam klasifikasi dan termasuk dalam 10 algoritma data mining terbaik [21]. Model ini banyak diminati karena output yang dihasilkan dapat lebih mudah dipahami oleh bahasa manusia. Algoritma ini pernah digunakan untuk klasifikasi diabetes tipe 2 dengan hasil akurasi 78,1768% [22]. Dalam penelitian ini tingkat akurasi decision tree adalah 95,96%. Gambar 12 merupakan tingkat akurasi confusion matrix dari decision tree yang dihasilkan dari penelitian ini. Gambar 13 merupakan model pohon keputusan yang terbentuk dari hasil klasifikasi.

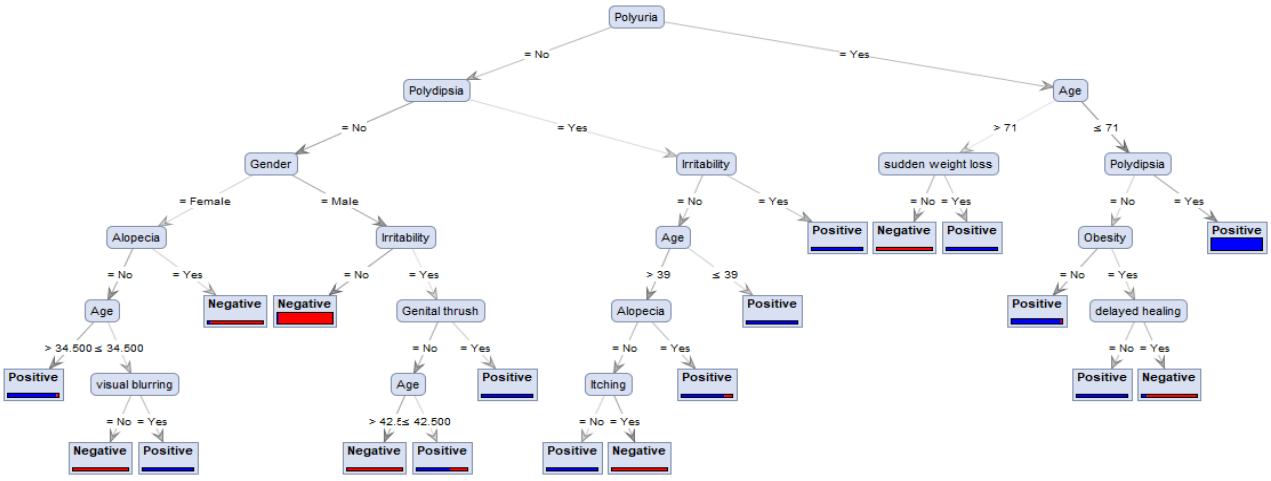
Multiclass Classification Performance Annotations

Table View Plot View

accuracy: 95.96% +/- 2.35% (mikro: 95.96%)

	true Positive	true Negative	class precision
pred. Positive	306	7	97.76%
pred. Negative	14	193	93.24%
class recall	95.62%	96.50%	

Gambar 12. Perhitungan decision tree dengan rapid miner



Gambar 13. Pohon keputusan hasil perhitungan dengan rapid miner

3.2 Pembahasan

Rangkuman hasil tingkat akurasi dari penelitian ini dapat dilihat pada tabel 3. Dalam tabel tersebut naïve bayes memiliki tingkat akurasi terkecil yaitu 86,92% dikarenakan algoritma ini cocok untuk tipe data nominal. Sedangkan di dalam

ISSN 2774-3639 (Media Online)
Vol 3, No 5, Agustus 2023 | Hal 343-350
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v3i5.280

dataset terdapat satu atribut dengan tipe data integer yaitu atribut usia (age). Dengan variasi data yang banyak dalam atribut usia membuat probabilitas naïve bayes menjadi lebih banyak dan dapat menyebabkan hasil klasifikasi kurang optimal. Pada KNN dan decision tree atribut integer tidak terlalu berpengaruh. Dalam perhitungan KNN atribut data numeric ataupun nominal tidak terlalu berpengaruh dalam hasil klasifikasi. Yang paling mempengaruhi hasil klasifikasi KNN adalah pembobotan atribut. Dalam penelitian ini semua atribut diberikan pembobotan yang sama. Dalam perhitungan decision tree nilai gain untuk setiap atribut dihitung untuk menentukan simpul atau cabang pertama dan diulang sampai dengan semua atribut menjadi cabang. Ini membuat decision tree kuat terhadap tipe data numeric ataupun nominal. Dikarenakan atribut numeric yang memiliki varian yang banyak pastinya akan memiliki nilai gain yang tidak terlalu tinggi sehingga tidak akan dijadikan simpul pertama. Secara lebih terperinci perbandingan tingkat akurasi dari algoritma naïve bayes, knn dan decision tree dapat dilihat pada tabel 3 berikut.

Tabel 3. Perbandingan tingkat akurasi algoritma

Algoritma	Akurasi	presisi
Naïve Bayes	86,92%	80,91%
KNN	92,5%	87,04%
Decission tree	95,96%	93,27%

4. KESIMPULAN

Dalam klasifikasi early stage diabetes risk prediction dataset decission tree memiliki tingkat akurasi terbaik yaitu 95,96%. Sedangkan KNN hanya memperoleh tingkat akurasi sebesar 92,5% dan naïve bayes dengan tingkat akurasi 86,92%. Penelitian berikutnya dapat menerapkan pre processing data guna meningkatkan akurasi dari decision tree untuk klasifikasi early stage diabetes risk prediction dataset.

REFERENCES

[1] WHO, “Diabetes,” 2023. https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Jul. 25, 2023).
[2] C. J. Ejiyi et al., “A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms,” Healthc. Anal., vol. 3, no. December 2022, p. 100166, 2023, doi: 10.1016/j.health.2023.100166.
[3] O. Maimoon and L. Rokach, Data Mining and Knowledge Discovery Handbook, vol. 40, no. 6. Springer, 2010.
[4] J. Han and M. Kamber, “Data Mining: Concepts and Techniques Second Edition,” vol. 40, no. 6, p. 9823, Mar. 2006, doi: 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.
[5] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition. Elsevier, 2011.
[6] X. Wu et al., Top 10 algorithms in data mining, vol. 14, no. 1. 2007.
[7] ikhsan wisnuadji Gamadarenda and I. Waspada, “Implementasi Data Mining Untuk Deteksi Penyakit Ginjal Kronis (Pgk) Menggunakan K-Nearest Neighbor (Knn) Dengan Backward Elimination,” vol. 7, no. 2, pp. 417–426, 2018, doi: 10.25126/jtiik.202071896.
[8] M. F. Kurniawan and Ivandari, “Komparasi Algoritma Data Mining untuk Klasifikasi Kanker Payudara,” IC Tech, vol. I April 20, pp. 1–8, 2017.
[9] G. Aguilera-Venegas, A. López-Molina, G. Rojo-Martínez, and J. L. Galán-García, “Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus,” J. Comput. Appl. Math., vol. 427, p. 115115, 2023, doi: 10.1016/j.cam.2023.115115.
[10] S. Kumari, D. Kumar, and M. Mittal, “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,” Int. J. Cogn. Comput. Eng., vol. 2, no. January, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.
[11] C. Carpinteiro, J. Lopes, A. Abelha, and M. F. Santos, “A Comparative Study of Classification Algorithms for Early Detection of Diabetes,” Procedia Comput. Sci., vol. 220, pp. 868–873, 2023, doi: 10.1016/j.procs.2023.03.117.
[12] S. Diabetes and B. Hospital in Sylhet, “Early stage diabetes risk prediction dataset,” 2020. https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset.
[13] Ian H Witten. Eibe Frank. Mark A Hall, Data Mining 3rd. 2011.
[14] Ivandari and M. A. Al Karomi, “Classification of Covid-19 Survillance Datasets using the Decision Tree Algorithm,” Jaict, vol. 6, no. 1, pp. 44–49, 2021, [Online]. Available: https://jurnal.polines.ac.id/index.php/jaict/article/view/2896.
[15] Ivandari and M. A. Al Karomi, “Algoritma K-NN untuk klasifikasi dataset Covid-19 survillance,” IC Tech, vol. 16, no. 1, pp. 12–15, 2021, [Online]. Available: https://ejournal.stmik-wp.ac.id/index.php/ictech/article/view/137.
[16] F. Gorunescu, Data Mining: Concepts; Models and Techniques. Springer, 2011.
[17] M. A. Alkaromi, “Komparasi Algoritma Klasifikasi untuk dataset iris dengan rapid miner,” IC Tech, vol. XI, no. 2, 2014.
[18] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” vol. I, 1967.
[19] Kusrini and L. E. Taufiq, Algoritma Data Mining. Yogyakarta: Andi Offset, 2009.
[20] Ivandari, “Improved Performance Algorithm K-Nearest Neighbor Classification in High Dimension Data,” IC Tech, vol. IX-April 2, pp. 5–9, 2014.
[21] V. K. Xindong Wu, The Top Ten Algorithm in Data Mining. 2009.
[22] A. A. Aljarullah, “Decision Tree Discovery for the Diagnosis of Type II Diabetes,” in International Conference on Innovations in Information Technology, 2011, pp. 303–307.