Infotek : Jurnal Informatika dan Teknologi
Vol. 4 No. 1, Januari 2021
Hal. 63-69

e-ISSN 2614-8773

DOI : 10.29408/jit.v4i1.2994 Link : https://dx.doi.org/10.29408/jit.v4i1.2994

Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes

Baiq Andriska Candra Permana[1] , Intan Komala Dewi[2]
[1]Program Studi Teknik Informatika, Universitas Hamzanwadi
[2]Program studi Teknik Komputer, Universitas Hamzanwadi
*andriska.cp@gmail.com

Abstrak

Diabetes merupakan suatu kelompok penyakit metabolik yang ditunjukan dengan terjadinya hiperglikemian yang disebabkan oleh adanya kelainan pada sekresi insulin di dalam tubuh. Banyak kematian yang diakibatkan oleh diabetes, bahkan jika tidak segera di tanggulangi diabetes dapat menyebabkan kerusakan pada bagian organ tubuh yang lain seperti kebutaan, store, gagal jantung bahkan gagal ginjal. Diperlukan suatu metode terbaik dalam melakukan klasifikasi penyakit diabetes dengan tujuan untuk dapat mengetahui secara dini seseorang mengalami diabetes. Penelitian terkait klasifikasi penyakit diabetes dengan menggunakan beberapa metode kalsifikasi sudah pernah dilakukan sebelumnya. Pada penelitian ini dilakukan komparasi dua metode klasifikasi yaitu decision tree dan naïve bayes, pengukuran metode dilakukan melalui cross validasi. Hasil yang diperoleh dari penelitian ini merupakan algoritma terbaik diantara kedua algoritma untuk menentukan penderita penyakit diabetes.

Kata Kunci : diabetes,klasifikasi,datamining,decision tree, naïve bayes

Abstract

Diabetes is a group of metabolic diseases which is indicated by the occurrence of hyperglycemia caused by abnormalities in insulin secretion in the body. Many deaths are caused by diabetes, if this disease is not treated immediately, diabetes can cause damage to other organs such as blindness, stores, heart problem and even kidney problem. A best method is needed in classifying diabetes in order to detect diabetes early. Research related to the classification of diabetes using several calcification methods has been done before. In this study, two classification methods were compared, namely decision tree and naïve Bayes. Measurement methods were carried out through cross validation. The results obtained from this study are the best algorithms among the two algorithms to determine diabetes sufferers

Keyword : diabetes, classification , data mining, decision tree, naïve bayes

1. Pendahuluan

Diabetes merupakan salah suatu penyakit yang menduduki peringkat teratas untuk golongan penyakit tidak menular yang menjadi penyebab kematian terbanyak di seluruh dunia. Menurut situr resmi WHO, 422 juta jiwa di seluruh dunia mengalami diabetes dan kematian yang disebabkan oleh diabetes mencapai 1.6 juta jiwa setiap tahunnya.

Diabetes atau yang lebih dikenal dengan Diabetes Militus (DM) merupakan suatu penyakit dimana kandungan guala yang terdapat didalam darah tidak dapat diolah dengan baik oleh tubuh [1]

Infotek : Jurnal Informatika dan Teknologi
Vol. 4 No. 1, Januari 2021
Hal. 63-69

e-ISSN 2614-8773

DOI : 10.29408/jit.v4i1.2994 Link : https://dx.doi.org/10.29408/jit.v4i1.2994

Dengan melihat banyaknya jumlah kasus diabetes, maka diperlukan tindakan awal untuk penanganan dini penyakit diabetes dengan melakukan predisksi.

Prediksi terhadap penderita diabetes dapat diperoleh melalui kumpulan beberapa data pasien penderita diabetes yang tersimpan dalam basis data, kemudian diolah dengan suau pola tertentu sehingga hasil nya dapat dugunakan untuk diagnose awal diabetes [2].

Sudah banyak penelitian yang dilakukan untuk melakukan prediksi diantaranya dengan decision tree maupun naïve bayes. Dalam penelitian ini akan dilakukan komparasi kedua algoritma tersebut untuk mendapatkan model dengan tingkat akurasi terbaik sehingga hasil prediksi bisa lebih akurat.

2. Tinjauan Pustaka

2.1 Penelitian Terkait

Penelitian dilakukan Hendri M, Jajang JP, dan Agung (2019) yaitu melakukan komparasi algoritma NN dan naïve bayes untuk prediksi penyakit jantung. Penelitian yang dilakukan menggunakan dataset dengan julah atribut sebanyak 14 atribut dan hasil yang diperoleh nilai klasifikasi AUC untuk NN sebesar 0.601% dan naïve bayaes sebesar 0.577 % dimana dapat dikatakan hasil kurang baik dan diperlukan penelitian lebih lanjut [3].

Penelitian oleh Abdul Rohman (2016), pada penelitian yang dilakukan yaitu melakukan komparasi metode kalsifikasi data mining untuk prediksi penyakit jantung dengan menggunakan algoritma NN, C4.5 dan K-nearest neighboard. Nilai akurasi yang terbaik ditunjukan oleh NN dengan nilai akurasi 86.06%, C4.5 dengan nilai akurasi sebesar 82.92% dan K-nearest neighboard nilai akurasi 77.58%. Sementara nilai AUC untuk NN tertinggi yaitu 0.913 [4].

2.2 Landasan Teori

1. Data Mining

Data mining merupakan suatu istilah yang sering digunakan untuk mendapatkan atau mencari suatu pengetahuan dalam suatu database [5].

Data mining dapat melakukan beberapa tugas meliputi : deskripsi, estimasi, prediksi , klasifikasi, clustering dan asosiasi (Larose,2005).

Beberapa model klasifikasi yang paling sering digunakan diantaranya decision tree, naïve bayes, neural network,genetic algorithms, K-nearst neighbor, support vector machine.

2. Decision tree

Algoritma C4.5 atau yang biasa di sebut sebagai decision tree memiliki training sample berupa sekumpulan data yang nantinya akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya [6]. Secara umum algoritma C4.5

Infotek : Jurnal Informatika dan Teknologi
Vol. 4 No. 1, Januari 2021
Hal. 63-69

e-ISSN 2614-8773

DOI : 10.29408/jit.v4i1.2994 Link : https://dx.doi.org/10.29408/jit.v4i1.2994

adalah untuk membangun pohon keputusan sebagai berikut :

a. Pilih atribut sebagai akar
b. Buat cabang untuk setiap nilai
c. Bagi kasus dalam cabang
d. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cangan memiliki kelas yang sama.

Rumus menghitung nilai entropy menggunakan persamaan [7] :

Entropy (S)= $\sum_{i=1}^{n} -\text{pi} \log_2 \text{pi}$ .....(1)

Keterangan :

S = himpunan kasus

n = jumlah partisi atribut A

Pi = proporsi Si terhadap S

|Si| = jumlah kasus pada partisi ke i

|S| = jumlah kasus dalam S

A= atribut

Rumus untuk mencari nilai gain :

Gaint (S,A) = $\sum_{f=1}^{n} \frac{|Si|}{|S|}$+ Entropy (Si) ..... (2)

3. Naïve Bayes

Naive bayes merupakan teori Bayesian dimana digunakan sebagai alat pengambilan keputusan dari suatu informasi [8] . Metode ini cukup banyak digunakan untuk hal yang terkait dengan diagnose secara statistik terkait probabilistic serta kemungkinan suatu penyakit dan gejala terkait. Teori bayes memiliki rumus berikut [9] :

$$P(H|E) = \frac{P(E|H)xP(H)}{P(E)} \quad \text{.....(3)}$$

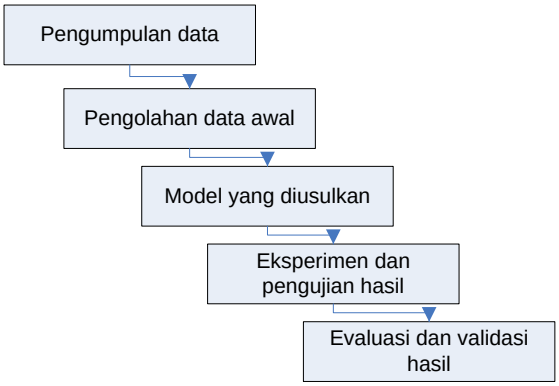
3. Metode Penelitian

Data yang digunakan pada penelitian ini berasal dari data para penderita penyakit diabetes yang diperoleh dari repository dengan alamat website : https://www.kaggle.com/ishandutta/early-stage-diabetes-risk-prediction-dataset. Data pada dataset tersebut merupakan data para penderita diabetes yang sudah melewati pemeriksaan oleh dokter.

3.1 Metode Pengumpulan Data Awal

Dataset yang digunakan merupakan dataset primer terdiri atas 520 data pasien dan 17 atribut meliputi (age, gender, polyuria,polydipsia,sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity,class).

Tahapan yang akan dilakukan dalam penelitian ini :



Gambar 1. Tahapan Penelitian

Infotek : Jurnal Informatika dan Teknologi
Vol. 4 No. 1, Januari 2021
Hal. 63-69

e-ISSN 2614-8773

DOI : 10.29408/jit.v4i1.2994 Link : https://dx.doi.org/10.29408/jit.v4i1.2994

3.2 Pengolahan data awal

Guna mendapatkan data dengan kualitas yang baik beberapa tahapan dapat dilakukan :

1. Data validation , digunakan untuk menghilangkan data yang dianggap tidak lengkap (missing value)
2. Data integration and transformation, digunakan untuk meningkatkan akurasi dan efisiensi algoritma.
3. Data size reduction and dicritization untuk memperoleh dataset dengan jumlah atribut dan record yang lebih sedikit tapi bersifat informatif.

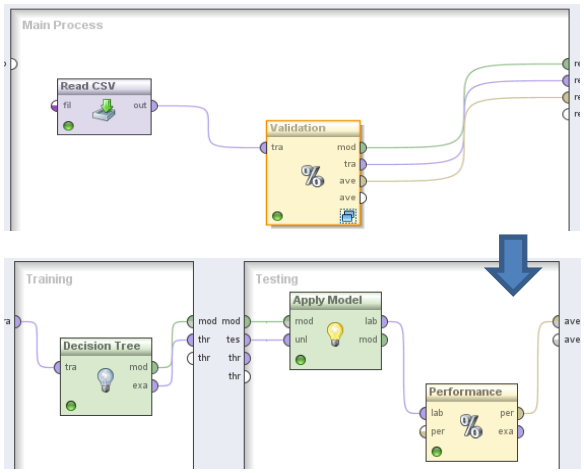
3.3 Metode Yang Diusulkan

Pada penelitian ini komparasi algoritma yang akan dilakukan adalah membandingkan algoritma klasifikasi decision tree dan naïve bayes dengan menggunakan cross validation untuk melakukan pengujian model, selanjutnya evaluasi yang dilakukan dengan confusion matrix untuk mendapatkan algoritma yang terbaik dalam memprediksi penyakit diabetes.

4. Hasil dan Pembahasan

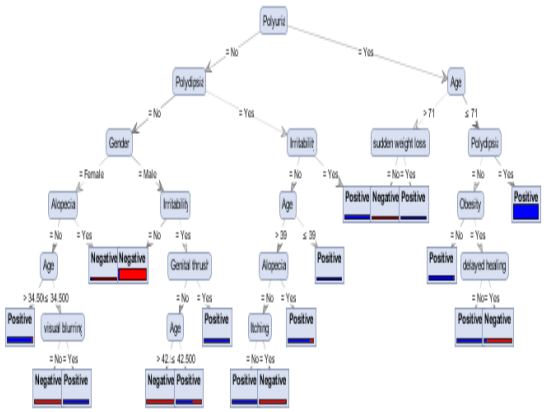
4.1 Eksperimen dan Pengujian Metode Algoritma Decision Tree

Pengolahan data menggunakan algoritma Decision Tree menggunakan tools rapidminer dengan 10 validasi ditunjukan oleh gambar berikut



Gambar 2. pengujian model Decision Tree menggunakan Rapid miner

Diperoleh pohon keputusan sebagaimana yang ditujukan pada gambar dengan pola – pola tertentu sebagai mana ditunjukan oleh gambar berikut :



Gambar 3. Pohon Keputusan

Untuk mendapatkan root pada decision tree dilakukan dengan terlebih dahulu menentukan nilai entropy . Tabel berikut menunjukan perhitungan nilai entropy dengan melihat jumlah kasus positif dan kasus negative pasien penderita diabetes :

Infotek : Jurnal Informatika dan Teknologi
Vol. 4 No. 1, Januari 2021
Hal. 63-69

e-ISSN 2614-8773

DOI : 10.29408/jit.v4i1.2994 **Link : https://dx.doi.org/10.29408/jit.v4i1.2994**

		All Case	Positive	Negative	Entropy	Gain
Total		520	320	200	**0.9612**	
POLYURIA	YES		243	15	**0.509**	**0.404**
	NO		77	185	**0.634**	
GENITAL THRUSH	YES		83	33	**0.843**	**0.415**
	NO		237	167	**0.073**	
VISUAL BLURRING	YES		175	58	**0.923**	**0.128**
	NO		145	142	**0.687**	
ITCHING	YES		154	99	**1.031**	**0.050**
	NO		166	101	**0.721**	
.....	...		...	...	**...**	**...**
DELAYED HEALING	YES		153	86	**1.018**	**0.074**

Gambar 4. Tabel entropy dan gain

dengan formula (1) :

Entropy (S)= $\sum_{i=1}^{n} -\text{pi} \log_2 \text{pi}$

$$\left(-\frac{320}{520}\log_2\frac{320}{520}+\left(-\frac{200}{520}\log_2\frac{200}{520}\right)\right)=0.96123660$$

$$-\frac{\log\left(\frac{243}{320}\frac{243}{320}\right)}{\log(2)}-\frac{\log\left(\frac{3}{64}\frac{3}{64}\right)}{\log(2)}$$

$$\left(-\frac{243}{320}\log_2\frac{243}{320}+\left(-\frac{15}{320}\log_2\frac{15}{320}\right)\right)=0.50851453$$

Poliyuria Negative :

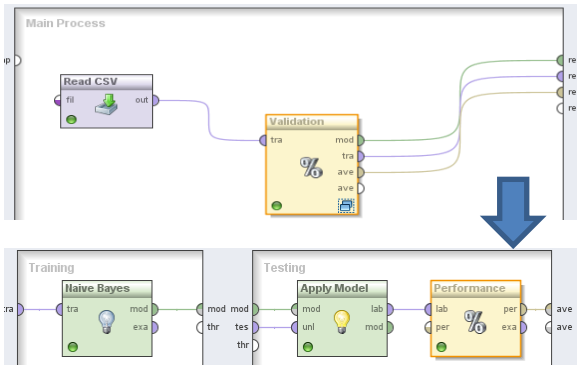
$$\left(-\frac{77}{200}\log_2\frac{77}{200}+\left(-\frac{185}{200}\log_2\frac{185}{200}\right)\right)$$
$$=0.63421093$$

Poliyuria gain :

$$\left(\begin{array}{c}0.96123660-\left(\left(\frac{320}{520}\right)*0.50851453\right)+\\ \left(\left(\frac{200}{520}\right)*0.63421093\right)\end{array}\right)$$
$$=0.404$$

4.2 Eksperimen dan pengujian model Naïve Bayes

Pengujian dengan menggunakan 10 validasi pada Naïve Bayes menggunakan tools rapid miner sebagai berikut :



Gambar 5. Pengujian model Naïve Bayes

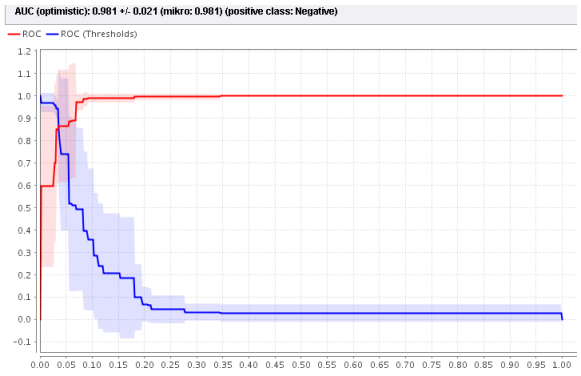
4.3 Evaluasi Dan Validasi Hasil

Setelah data diolah dengan menggunakan tools rapidminer diperoleh hasil berupa akurasi dari kedua model. untuk akurasi decision tree dan gambar 4.5 untuk akurasi menggunakan naïve bayes ditunjukan oleh gambar 4.7. Untuk pengujian tingkat akurasi sendiri menggunakan confusion matriks yang ditunjukan oleh gambar 4.6 untuk decision tree dan gambar 4.8 untuk naïve bayes.

accuracy: 95.58% +/- 2.86% (mikro: 95.58%)

	true Positive	true Negative	class precision
pred. Positive	306	9	97.14%
pred. Negative	14	191	93.17%
class recall	95.62%	95.50%	

Gambar 6. Nilai Akurasi Algoritma Decision Tree



Gambar 7. Curva AUC Algoritma Decision Tree

Infotek : Jurnal Informatika dan Teknologi
Vol. 4 No. 1, Januari 2021
Hal. 63-69

e-ISSN 2614-8773

DOI : 10.29408/jit.v4i1.2994 Link : https://dx.doi.org/10.29408/jit.v4i1.2994

Dari gambar 8 menunjukan bahwa dengan menggunakan Decision Tree , berdasarkan 520 data pasien diperoleh 306 pasien yang di prediksi diabetes dan benar diabetes, dan 9 orang yang di prediksi diabetes ternyata tidak diabetes. Sejumlah 191 orang di prediksi tidak diabetes dan benar, namun 14 yang di prediksi negative diabetes ternyata positif diabetes. Dengan demikian nilai akurasi yang di dapatkan dengan algoritma decision tree adalah 95,58%.

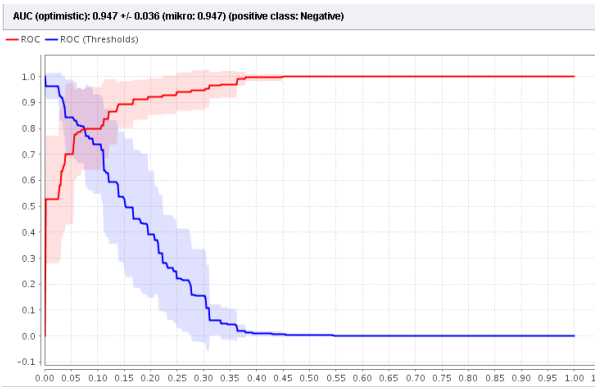
Gambar 9 menunjukan AUC (Area Under Curve) dimana garis horizontal menunjukan false positive dan garis vertical menunjukan false negative dari model decision tree sebesar 0,981.

Pengolahan dengan menggunakan algoritma naïve bayes menghasilkan nilai akurasi sebagaimana ditunjukan oleh gambar 4.7:

accuracy: 87.69% +/- 5.85% (mikro: 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. Nilai Akurasi Algoritma Naïve Bayes



Gambar 9. Curva AUC Algoritma Naïve bayes

Pada gambar 4.7 menunjukan bahwa dengan menggunakan algoritma naïve bayes , prediksi terhadap pasien penderita diabetes yang benar mengalami diabetes berjumlah 276 orang, sementara 20 orang yang di prediksi diabetes ternyata tidak diabetes. Sementara 44 orang yang diprediksi tidak diabetes ternyata diabetes , sementara 180 orang yang di prediksi tidak diabetes memang benar negative diabetes. Tingkat akurasi yang diperoleh dalam memprediksi menggunakan naïve bayes sebesar 87,69%.

Gambar 4.8 menunjukan nilai AUC untuk model naïve bayes sebesar 0.947.

5. Kesimpulan

Berdasarkan eksperimen yang telah dilakukan terhadap pasien penderita diabetes, diperoleh kesimpulan bahwa algoritma klasifikasi decision tree lebih baik dalam prediksi penyakit diabetes dengan nilai akurasi 95,58% dab nilai AUC 0,981 lebih tinggi dibandingkan naïve bayes dengan akurasi 87,69% dan nilai AUC 0,947.

Penelitian dapat dikembangkan dengan menggunakan metode optimasi sehingga nilai akurasi bisa lebih baik lagi.

Daftar Pustaka

[1] N. Hackworth *et al.*, “A Risk Factor Profile for Pre-diabetes: Biochemical, Behavioural,

e-ISSN 2614-8773
DOI : 10.29408/jit.v4i1.2994 Link : https://dx.doi.org/10.29408/jit.v4i1.2994

Psychosocial and Cultural Factors," *E-Journal Appl. Psychol.*, vol. 3, no. 2, 2007, doi: 10.7790/ejap.v3i2.89.

[2] P. Arsi and O. Somantri, "Deteksi Dini Penyakit Diabetes Menggunakan Algoritma Neural Network Berbasiskan Algoritma Genetika," *J. Inform. J. Pengemb. IT*, vol. 3, no. 3, pp. 290–294, 2018, doi: 10.30591/jpit.v3i3.1008.

[3] H. M. Nawawi, J. J. Purnama, and A. B. Hikmah, "Komparasi Algoritma Neural Network Dan Naïve Bayes Untuk Memprediksi Penyakit Jantung," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 189–194, 2019, doi: 10.33480/pilar.v15i2.669.

[4] A. Rohman, "Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Penyakit Jantung," *Neo Tek.*, vol. 2, no. 2, pp. 21–28,2017,doi:10.37760/neoteknika.v2i2.766.

[5] M. A. C, "Analisis Pola Penyebaran Penyakit dengan Menggunakan Algoritma C4.5" vol. 03, no. 479, pp. 3–7, 2018.

[6] Bq Andriska CP, I Gunawan, R. Ahmad, "Penggunaan Metode NN Untuk Mengukur Pengaruh Web Promosi Dan Faktor Harga Terhadap Penjualan Kain Tenun Oleh Pengrajin Di Pringgasela Lomobok Timur," vol. 2, no. 1, 2019.

[7] G. Syahputra, M. Kom, P. Studi, and M. Informatika, "Penerapan Algoritma C4 . 5 Dalam Analisa Kelayakan Penerima Bonus Tahunan Pegawai (Studi Kasus : PT . Multi Pratama Nauli Medan)," vol. 16, no. 2, 2015.

[8] Yupi Kuspandi Putra, Muhamad Sadali, "Perbandingan Algoritma Naïve Bayes dan Naïve Bayes Berbasis PSO untuk Analisis Kredit pada PT. BPR Syariah Paokmotong," vol. 8, no. 5, p. 55, 2019.

[9] N. T. Rahman, F. I. Komputer, U. Darwan, and A. Sampit, "Analisa Algoritma Decision Treedan Naïve Bayespada Pasien Penyakit Liver," vol. 10, no. 2, pp. 144–151, 2020.