KLASIFIKASI METODE *NAÏVE BAYES* DAN *DECISION TREE* ALGORITMA *ITERATIVE DICHOTOMISER THREE (ID3)* PADA PENYAKIT DIABETES

SKRIPSI

Sebagai Salah Satu Syarat untuk Memproleh Gelar
Sarjana Sains Bidang Matematika

Oleh:

RAHMA PUTRI DEWI
08011381924108

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SRIWIJAYA
2023

HALAMAN PENGESAHAN

KLASIFIKASI METODE *NAÏVE BAYES* DAN *DECISION TREE* ALGORITMA *ITERATIVE DICHOTOMISER THREE (ID3)* PADA PENYAKIT DIABETES

SKRIPSI

Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Sains Bidang Studi Metematika

Oleh :

RAHMA PUTRI DEWI
NIM. 08011381924108

Indralaya, April 2023

Pembimbing Pembantu	Pembimbing Utama
Des Alwine Zayanti, S.Si., M.Si	**Endang Sri Kresnawati, S.Si., M.Si**
NIP. 197012041998022001	**NIP. 197702082002122003**

Mengetahui,
Ketua Jurusan Matematika

Drs. Sugandi Yahdin, M.M
NIP. 195807271986031003

PERNYATAAN KEASLIAN KARYA ILMIAH

Yang bertanda tangan dibawah ini:

Nama Mahasiswa	: Rahma Putri Dewi
NIM	: 08011381924108
Fakultas/Jurusan	: Matematika dan Ilmu Pengetahuan Alam/Matematika

Menyatakan bahwa skripsi ini adalah hasil karya saya sendiri dan karya ilmiah ini belum pernah diajukan sebagai pemenuhan persyaratan untuk memperoleh gelar kesarjanaan strata satu (S1) dari Universitas Sriwijaya maupun perguruan tinggi lain.

Semua informasi yang dimuat dalam skripsi ini yang berasal dari penulis lain baik yang dipublikasikan atau tidak telah diberikan penghargaan dengan mengutip nama sumber penulis secara benar. Semua isi dari skripsi ini sepenuhnya menjadi tanggung jawab saya sebagai penulis.

Demikianlah surat pernyataan ini saya buat dengan sebenarnya.

Indralaya, April 2023

Penulis

Rahma Putri Dewi

NIM. 08011381924108

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Puji syukur kehadiran Allah SWT atas nikmat dan karunia-nya sehingga penulis dapat menyelesaikan skripsi yang berjudul **"Klasifikasi Metode *Naïve Bayes* Dan *Decision Tree* Algoritma *Iterative Dichotomizer Three (Id3)* pada Penyakit Diabetes"** dimana tugas akhir ini dapat berjalan dengan baik. Shalawat serta salam semoga selalu tercurah kepada junjungan kita nabi Muhammad SAW beserta keluarga, sahabat, dan pengikutnya hingga akhir zaman. Skripsi ini merupakan salah satu syarat memperoleh gelar Sarjana Sains bidang studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya.

Dengan segala hormat penulis ucapkan terima kasih yang sebesar-besarnya kepada orang tua tercinta, yaitu **Bapak Abdullah Zikri,** dan **(Alm) Ibu Titik Supriati** serta ibu sambung saya **Ibu Utama Dewi** yang tak pernah lupa mendoakan yang terbaik untuk penulis, telah merawat dengan baik, membimbing, memberikan kasih sayang, dan selalu memberikan dukungan yang begitu besar terhadap penulis. Penulis juga menyampaikan ucapan terima kasih dan penghargaan kepada:

1. Bapak **Drs. Sugandi Yahdin, M.M,** selaku ketua jurusan matematika fakultas matematika dan ilmu pengetahuan alam universitas sriwijaya yang telah membimbing serta mengarahkan penulis untuk urusan akademik selama di jurusan matematika fakultas matematika ilmu dan ilmu pengetahuan alam universitas sriwijaya.

2. Ibu **Dr. Dian Cahyawati Sukanda, M.Si** selaku sekertaris jurusan matematika fakultas matematika dan ilmu pengetahuan alam universitas sriwijaya yang telah membimbing serta mengarahkan penulis untuk urusan akademik selama di jurusan matematika fakultas matematika ilmu dan ilmu pengetahuan alam universitas sriwijaya.
3. Ibu **Ending Sri Kresnawati, M.Si** selaku dosen pembimbing utama yang telah bersedia meluangkan waktu, tenaga, pikiran untuk memberikan bimbingan dan pengarahan dengan penuh perhatian dan kesabaran sehingga skripsi ini dapat diselesaikan dengan baik.
4. Ibu **Des Alwine Zayanti, M.Si** selaku dosen pembimbing kedua yang telah bersedia meluangkan waktu, tenaga, pikiran serta nasehat dalam membimbing penulis dalam penulisan skripsi sehingga dapat diselesaikan dengan baik.
5. Ibu **Dr. Yulia Resti, M.Si dan bapak Drs. Ali Amran, M.T** selaku dosen pembahas yang telah bersedia meluangkan waktu, tanggapan serta kritik dan saran yang sangat bermanfaat dalam menyelesaikan skripsi ini.
6. Seluruh Dosen di Jurusan Matematika Fakultas matematika dan ilmu pengetahuan alam universitas sriwijaya yang mana telah memberikan bimbingan, nasehat, serta ilmu yang bermanfaat bagi penulis selama menjalankan perkuliahan.
7. Bapak **Irwansyah** selaku admin dan ibu khamidah serta pegawai tata usaha jurusan matematika fakultas ilmu pengetahuan alam universitas sriwijaya yang telah membantu selama perkuliahan.

8. Kakakku **Anita Mulya fuji astuti** dan adikku **Dimas, Adit, Lolita, Nayla** serta keluargaku tersayang terimakasih yang telah mendoakan, membantu, selama perkuliahan.
9. Sepupuku **Nindy** yang selalu mendukung serta membantu dalam pembuatan skripsi.
10. Teman-teman seperjuanganku dalam skripsi **Andini, Anggraini, Siwi, Tasya, Vira, Zahra, Gaya, Miranda, Natalia, Lynda, Yunia** yang selalu mendukung dan membantu selama perkuliahan.
11. Semua pihak yang tidak dapat penulis sebutkan satu persatu. Saya ucapkan terima kasih semoga amal baik semua pihak mendapat balasan yang berlipat ganda dari Allah SWT.

Indralaya, April 2023

Penulis

CLASSIFICATION OF NAÏVE BAYES AND DECISION TREE METHODS ALGORITHM ITERATIVE DICHOTOMISER THREE (ID3) IN DIABETES

By :

Rahma Putri Dewi

08011381924108

ABSTRACT

Diabetes mellitus (DM) is a disorder of carbohydrate metabolism, in which glucose cannot be used properly, which can cause hyperglycemia. Hyperglycemia is a condition in which diabetes mellitus (DM) is not controlled in the body and results in very high blood glucose levels reaching more than 300 mg/dl. According to data from the International Diabetes Federation (IDF) for 2019, the number of diabetics in Indonesia will continue to increase from 10.7 million to 19.5 million in 2021. can see the level of accuracy of diabetes. In this study using secondary data obtained from Kaggle.com with a total dataset of 520 with 17 variables, the prediction of diabetes classification using the Naïve Bayes method and the Iterative Dichotomizer Three (Id3) algorithm. The results of this study obtained an accuracy rate of the Naïve Bayes method of 88.46% recall 86.46% and precision 94.32% while in the Iterative Dichotomizer Three (Id3) algorithm the accuracy was 95.51% recall 95.83% and precision 96, 84%.

Keywords: diabetes, Naïve Bayes, algoritma Iterative Dichotomiser Three (Id3) algorithm.

KLASIFIKASI METODE *NAÏVE BAYES* DAN *DECISION TREE* ALGORITMA *ITERATIVE DICHOTOMISER THREE (ID3)* PADA PENYAKIT DIABETES

Oleh :

Rahma Putri Dewi

08011381924108

ABSTRAK

Diabetes mellitus (DM) adalah kelainan metabolisme karbohidrat, dimana glukosa tidak dapat digunakan dengan baik, sehingga dapat menyebabkan keadaan hiperglikemia. Hiperglikemia adalah kondisi dimana diabetes mellitus (DM) pada tubuh tidak terkontrol dan mengakibatkan kadar glukosa darah sangat tinggi hingga mencapai lebih dari 300 mg/dl. Menurut data *International Diabetes Federation* (IDF) tahun 2019, jumlah penderita diabetes di indonesia terus bertambah dari 10,7 juta menjadi 19,5 juta pada tahun 2021. Oleh karena itu, dilakukan klasifikasi penyakit diabetes agar dapat memprediksi seseorang terkena diabetes atau tidak, serta dapat melihat tingkat akurasi penyakit diabetes. Dalam penelitian ini menggunakan data sekunder yang diperoleh dari *Kaggle.com* dengan jumlah dataset 520 dengan 17 variabel, prediksi klasifikasi penyakit diabetes menggunakan metode *Naïve Bayes* dan algoritma *Iterative Dichotomiser Three (Id3)*. Hasil dari penelitian ini memperoleh tingkat akurasi metode *Naïve Bayes* sebesar 88,46% recall 86,46% dan precision 94,32% sedangkan pada algoritma *Iterative Dichotomiser Three (Id3)* akurasi sebesar 95,51% recall 95,83% dan precision 96,84%.

Kata kunci : diabetes, *Naïve Bayes*, algoritma *Iterative Dichotomiser Three (Id3)*.

DAFTAR ISI

Halaman

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus (DM) adalah kelainan metabolisme karbohidrat, dimana glukosa tidak dapat digunakan dengan baik, sehingga dapat menyebabkan keadaan hiperglikemia. Hiperglikemia adalah kondisi dimana diabetes mellitus (DM) pada tubuh tidak terkontrol dan mengakibatkan kadar glukosa darah sangat tinggi hingga mencapai lebih dari 300 mg/dl. Terdapat 2 kategori utama diabetes melitus yaitu, diabetes tipe 1 dan diabetes tipe 2. Diabetes Tipe 1 ditandai dengan kurangnya produksi insulin yang merupakan kondisi kronis saat pankreas memproduksi insulin lebih sedikit atau tidak sama sekali. Diabetes tipe 2, disebabkan penggunaan insulin yang kurang efektif oleh tubuh. Diabetes tipe 2 merupakan 90% dari seluruh penderita diabetes (Ginting *et al*., 2022).

Menurut data *International Diabetes Federation* (IDF) tahun 2019, jumlah penderita diabetes di indonesia terus bertambah dari 10,7 juta menjadi 19,5 juta pada tahun 2021, ini menjadikan Indonesia sebagai negara penderita diabetes kelima terbanyak di dunia. Menurut laporan tahun 2021 oleh *International Diabetes Federation* (IDF), 19,5 juta orang Indonesia yang terkena diabetes berusia antara 20-79 tahun yang mengidap penyakit diabetes.

Machine Learning adalah aplikasi kecerdasan buatan dan ilmu komputer yang berfokus pada penggunaan data dan algoritma untuk meniru cara manusia belajar dan menjadi lebih akurat dari waktu ke waktu. *Machine Learning* secara efektif

digunakan dalam perawatan kesehatan, termasuk analisis citra medis, penemuan obat, diagnosis dan prognosis, skrining penyakit, dan prediksi wabah (Ginting *et al.,* 2022). Pada *Machine Learning* terdapat data mining dimana data mining merupakan suatu alat yang memungkinkan pengguna untuk mengakses data secara cepat dengan jumlah yang cukup besar (Wijaya & Dwiasnati, 2020).

Klasifikasi adalah proses mengidentifikasi objek ke dalam kategori, kelas, atau kelompok berdasarkan prosedur, definisi, dan karakteristik yang telah ditentukan. Klasifikasi bertujuan untuk menempatkan objek yang hanya dimiliki oleh salah satu kategori yang disebut kelas (Dwi *et al.*, 2022). Pada proses klasifikasi data dibagi menjadi dua bagian yaitu data latih *(training)* dan data tes *(testing)*. Pada data *training*, sebagian data yang sudah diketahui kelasnya kemudian data yang dicobakan untuk membentuk model perkiraan. Kemudian pada data *testing*, model yang sudah terbentuk diuji dengan sebagian data.

Beberapa metode yang sering digunakan diantaranya *Decision trees, Neural Networks, K-Nearst Neighbor dan Naïve Bayes*. Pada penelitian ini, metode yang digunakan untuk memprediksi klasifikasi tingkat seseorang terkena penyakit diabetes dengan gejala yang dialami yaitu menggunakan metode *Naïve Bayes* dan Algoritma *Interativ discoion Tree* (ID3).

Naïve Bayes merupakan klasifikasi yang menggunakan metode probabilistik dan statistik dengan memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya yang dikemukakan oleh ilmuan inggris *Thomas Bayes* (Widodo *et al*., 2021). Konsep dasar yang digunakan oleh *Naïve Bayes* adalah Teorema

Bayes, yaitu teorema dalam statistika untuk menghitung peluang dari satu kelas dari masing-masing kelompok atribut yang ada.

Berdasarkan penelitian terdahulu mengenai prediksi awal penyakit diabetes mellitus menggunakan algoritma *Naïve Bayes* yang dilakukan (Rumini & Nasruddin, 2021). Penelitian ini menggunakan data yang didapat pada situs *UCI Machine Learning,* pada *dataset* dilakukan dengan cara membagi data *training* dan data *testing* sebesar 80% dan 20%. Hasil dari penelitian tersebut memperoleh hasil akurasi yaitu 89%.

(Hafidh *et al*., 2021) Juga melakukan penelitian mengenai identifikasi ketunaan anak berkebutuhan khusus dengan algoritma *Iterative Dichotomiser Three (ID3)*. Proses klasifikasi terhadap anak berkebutuhan khusus dengan algoritma ID3 telah diuji performanya dengan *cross validation* dan menghasilkan tingkat akurasi sebesar 91,67%.

Berdasarkan hasil dari penelitian-penelitian terdahulu memperlihatkan bahwa metode *Naïve Bayes* dan Algoritma *Iterative Dichotomiser Three (ID3)* dapat melakukan klasifikasi dengan baik. Oleh karena itu, pada skripsi ini peneliti mencoba untuk mengklasifikasikan penyakit diabetes sedikit berbeda dengan penelitian sebelumnya, yaitu menggunakan lebih banyak dataset dengan partisi data *training* 70% dan data *testing* 30%, serta menambahkan metode yang digunakan. Peneliti menggunakan metode *Naïve Bayes* dan Algoritma *Iterative Dichotomiser Three (ID3)* untuk melihat perbandingan antara dua metode tersebut serta memprediksi penyakit diabetes berdasarkan gejala yang terjadi. Penelitian ini juga dapat memberikan informasi yang tepat kepada masyarakat mengenai penyakit

diabetes yang diderita agar masyarakat dapat menangani penyakit tersebut lebih lanjut.

1.2 Perumusan Masalah

Adapun rumusan masalah dalam penelitian ini sebagai berikut:

1. Bagaimana klasifikasi penyakit diabetes dengan menggunakan metode *Naïve Bayes dan* Algoritma *Iterative Dichotomiser Three (ID3)?*
2. Bagaimana hasil tingkat akurasi pada penyakit diabetes dengan menggunakan metode *Naïve Bayes* dan Algoritma *Iterative Dichotomiser Three (ID3)?*

1.3 Pembatasan Masalah

Adapun batasan masalah dalam penelitian ini sebagai berikut:

1. Dalam penelitian ini data yang digunakan sebanyak 520 data dengan 16 variabel independen dan 1 variabel dependen.
2. Variabel independen yang digunakan yaitu umur, jenis kelamin, *polyuria, polydipsia*, berat badan turun, keletihan, *polyphagia*, iritasi genital, penglihatan kabur, gatal, mudah marah, sembuh lambat, *partial paresis*, otot kaku, rambut rontok, obesitas.
3. Variabel dependen yang digunakan yaitu diabetes (kelas).
4. Data di partisi dengan menggunakan *split validation* 70% data *training* dan 30% data *testing.*
5. Persentase ketepatan klasifikasi pada penelitian ini dibatasi oleh nilai akurasi.

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini sebagai berikut:

1. Melakukan klasifikasi penyakit diabetes dengan menggunakan metode *Naïve Bayes* dan algoritma *Iterative Dichotomiser Three (ID3).*
2. Membandingkan tingkat akurasi metode *Naïve Bayes* dan Algoritma *Iterative Dichotomiser Three (ID3)* dalam klasifikasi penyakit diabetes.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini sebagai berikut:

1. Penelitian ini diharapkan bisa menjadi referensi untuk penelitian kasus yang berbeda dalam menggunakan metode *Naïve Bayes* Algoritma *Iterative Dichotomiser Three (ID3)*
2. Sebagai salah satu media pembelajaran dalam klasifikasi menggunakan metode *Naïve Bayes* dan algoritma *Iterative Dichotomiser Three (ID3)*

BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Data mining adalah suatu proses pengolahan data untuk menentukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan data mining ini dapat digunakan untuk mengambil keputusan di masa depan (Sulastri & Gufroni, 2017). *Machine Learning*, merupakan teknik untuk melakukan inferensi terhadap data dengan pendekatan matematis. *Machine Learning* adalah untuk membuat model (matematis) yang menggambarkan pola-pola data. *Machine Learning* bekerja apabila tersedia data sebagai input untuk dilakukan analisis terhadap kumpulan data besar sehingga menemukan pola tertentu.

Pada *Machine Learning* terdapat dua istilah penting dalam pembangunan model *Machine Learning* yaitu *training* dan *testing*. *Training* adalah proses konstruksi model dan *testing* adalah proses menguji kinerja model pembelajaran (Putra, 2020). *Machine Learning* menggunakan teknik untuk menangani data besar dengan cara yang cerdas untuk memberikan hasil yang tepat. Berdasarkan teknik pembelajarannya, tipe-tipe *machine learning* dapat dibedakan menjadi *supervised learning, unsupervised learning, semi supervised learning* dan *reinforcement learning. Supervised learning* merupakan salah satu teknik machine learning yang menggunakan *dataset* (data *training*) yang sudah berlabel untuk melakukan pembelajaran pada mesin, sehingga mesin mampu mengidentifikasi label input dengan menggunakan fitur yang dimiliki untuk selanjutnya melakukan prediksi maupun klasifikasi, metode klasifikasi yang termasuk kedalam teknik *supervised*

learning diantaranya *Decision tree, K-Nearest Neighbor (KNN), Naive Bayes, Regresi,* dan *Super Vector Machine* (Retnoningsih & Pramudita, 2020).

2.2 Metode Klasifikasi

Klasifikasi (*Classification)* merupakan teknik yang digunakan untuk mengambil informasi penting dan relevan tentang data. Teknik ini juga digunakan untuk mengklasifikasikan data yang berbeda di kelas yang berbeda. Secara sederhana dalam teknik klasifikasi akan menggunakan metode klasifikasi untuk memutuskan bagaimana mengklasifikasikan data baru yang akan digunakan (Budiyantara *et al.*, 2021).

2.3 Metode *Naïve Bayes*

Naive Bayes adalah pengklasifikasi probabilistik sederhana yang menghitung berbagai kemungkinan dengan menjumlahkan frekuensi dan kombinasi nilai dalam kumpulan data yang ada. Definisi lain menyebutkan bahwa Naive Bayes adalah classifier yang menggunakan metode probabilistik dan statistik yang diperkenalkan oleh ilmuwan Inggris Thomas Bayes, yaitu untuk memprediksi kemungkinan masa depan berdasarkan pengalaman masa lalu. Naive Bayes didasarkan pada asumsi penyederhanaan yang diberi nilai awal, nilai atribut secara kondisional independen. Dengan kata lain, mengingat nilai output, probabilitas observasi kolektif adalah produk dari probabilitas individu. Kelebihan Naive Bayes adalah metode ini hanya membutuhkan sedikit data latih untuk menentukan estimasi parameter yang dibutuhkan dalam proses klasifikasi. *Naïve Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan (Widodo *et al*., 2021).

Persamaan dari teorema bayes adalah:

$$P(Y|X_1, \dots, X_n) = \frac{P(Y)P(X_1, \dots, X_n|Y)}{P(X_1, \dots, X_n)} \quad 2.1$$

Di mana variabel $X_1, \dots, X_n$ merupakan variabel independen untuk melakukan proses klasifikasi. Sementara variabel Y merupakan variabel dependen untuk mempresentasikan kelas. $P(Y|X_1, \dots, X_n)$ merupakan peluang masuknya sampel variabel independen ke dalam variabel dependen yang disebut *posterior*. P(Y) merupakan peluang kemunculan pada variabel dependen (Y) yang disebut *prior*. $P(X_1, \dots, X_n|Y)$ merupakan peluang munculnya variabel dependen yang masuk ke dalam variabel independen yang disebut dengan *likelihood*. Kemudian $P(X_1, \dots, X_n)$ merupakan peluang munculnya variabel independen secara keseluruhan yang disebut *evidence.* Persamaan diatas dapat ditulis juga sebagai berikut:

$$posterior = \frac{prior \times likelihood}{evidence} \quad 2.2$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari *posterior* tersebut nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan ke kelas mana suatu sampel akan diklasifikasikan (Rosandy, 2016).

2.4 Algoritma *Iterative Dichotomiser Three (ID3)*

Algoritma *Iterative Dichotomiser Three (ID3)* merupakan bagian dari metode *Decision tree* (pohon keputusan). *Decision tree* merupakan *flowchart* seperti struktur pohon, dimana setiap titik pohon merupakan variabel yang telah diuji. Bagian awal dari pohon keputusan disebut titik akar (*root node*) sedangkan setiap

cabang (*branch node*) dari *decision tree* merupakan pembagian berdasarkan hasil uji, dan titik akhir disebut node daun (*leaf node*) yang mana merupakan pembagian kelas yang dihasilkanan (Nugroho *et al.*, 2007). Algoritma *Iterative Dichotomiser Three (ID3)* pertama kali diperkenalkan dan dikembangkan oleh *J. Ross Quinlan.* ID3 akan membentuk pohon keputusan yang dimulai dari atas ke bawah (*top down*). Input dari ID3 yaitu sebuah database dengan beberapa variabel yang juga dikenal sebagai atribut (Hikmatulloh *et al.*, 2019).

Proses klasifikasi dilakukan dari node paling atas yaitu akar pohon *(root node).* Dilanjutkan ke bawah melalui cabang-cabang sampai dihasilkan node daun (*leaf node*) dimana node daun ini menunjukan hasil akhir klasifikasi. Sebuah objek yang diklasifikasikan dalam pohon harus dites nilai entropinya. Dari nilai entropi tersebut kemudian dihitung nilai information gain (IG) masing-masing variabel independen terhadap variabel dependen. IG merupakan nilai rata-rata entropi pada semua variabel (Giovani *et al.*, 2011). Information gain dapat dimaknai sebagai pengurangan entropy karena melakukan percabangan. Tujuan dari menghitung nilai information gain adalah untuk memilih variabel independen yang akan dijadikan cabang pada pembentukan pohon keputusan. Variabel yang memiliki nilai information gain tertinggi akan dipilih menjadi variabel uji untuk dijadikan cabang pohon. Berikut ini rumus entropy dan informasion gain yang diperlukan mengkelaskan sampel adalah :

$$entropy(S) = -\sum_{i=1}^{n} p_i \ log_2 \ p_i \qquad 2.3$$

$S =$ himpunan kasus (sampel)

$n =$ jumlah partisi S

p_i =peluang yang didapat dari jumlah kelas dibagi total kasus

$$information\ gain\ (S, A) = -\sum_{i=1}^{n} \frac{S_i}{S}\ entropy\ (S_i) \quad 2.4$$

n = jumlah partisi variabel independen

S_i = jumlah kasus pada partisi ke-i

A = variabel independent

(Laila, 2021).

2.5 *Preprocessing* Data

Preprocessing data dilakukan untuk memproses data mentah yang diambil kemudian diubah menjadi data dengan format yang dapat dipahami dan dianalisis oleh *Machine Learning* (Made *et al*., 2020). Diskritisasi merupakan salah satu teknik untuk merubah suatu nilai kontinu kedalam bentuk diskrit. Teknik ini dilakukan sebagai penyesuaian nilai kontinu yang kemungkinan muncul dalam fitur *dataset* sehingga akan mempengaruhi proses klasifikasi dengan menggunakan metode *Naïve Bayes* (Wirawan & Eksistyanto, 2015).

2.6 *Confusion Matrix*

Confusion Matrix adalah salah satu metode yang digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada umumnya metode ini memberikan informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sebuah sistem dengan hasil klasifikasi yang seharusnya (Motif *et al*., n.d.). *Confusion matrix* akan digunakan dalam mengelompokkan data klasifikasi kedalam empat bagian, dan akan digunakan untuk menghitung besar akurasi pengujian.

Table 2.1 *Confusion Matrix*

Actual Class	***Predicted Class***	
	Yes	***No***
Yes	TP	FN
No	FP	TN

Keterangan:

TN = True Negative (jumlah data kelas negatif yang diklasifikasikan sebagai kelas negatif)

FP = False Positive (jumlah data kelas negatif yang diklasifikasikan sebagai kelas positif)

FN = False Negative (jumlah data kelas positif yang diklasifikasikan sebagai kelas negatif)

TP = True Positive (jumlah data kelas positif yang diklasifikasikan sebagai kelas positif)

Ukuran tingkat ketepatan proses klasifikasi yang dihasilkan pada tabel 2.1 *confusion matrix* sebagai berikut:

$$Akurasi \quad = \frac{TP + TP}{TP + FP + TN + FN} 100\% \qquad 2.5$$

Akurasi adalah hasil evaluasi kenerja model dalam proses klasifikasi.

$$Recall \quad = \frac{TP}{TP + FN} 100\% \qquad 2.6$$

Recall adalah perbandingan antara *true* positif (TP) dengan banyaknya data yang sebenarnya positif.

$$Precision = \frac{TP}{TP + FP} 100\% \qquad 2.7$$

Precisison adalah perbandingan antara *true* positif (TP) dengan banyaknya data yang diprediksi positif.

2.7 Diabetes

Diabetes mellitus merupakan penyakit metabolik yang ditandai dengan meningkatnya gula darah yang disebabkan terganggunya hormone insulin, hormone insulin ini digunakan untuk memelihara kondisi tubuh agar relatif stabil dengan cara penurunan kadar gula darah (Widiyoga *et al.*, 2020). Diabetes merupakan salah satu penyakit dengan kematian yang cukup tinggi. Berdasarkan data pada kasus penyakit yang ada di Indonesia dari Litbang Kemenkes terdapat jumlah kasus baru Penyakit Diabetes sebesar 56,82%. Penyakit diabetes merupakan penyakit yang berlangsung cukup lama atau kronis serta ditandai dengan kadar gula (glukosa) darah yang tinggi atau di atas nilai normal. Glukosa yang menumpuk di dalam darah akibat tidak diserap sel tubuh dengan baik dapat mengakibatkan berbagai gangguan organ tubuh (Dwi *et al.*, 2022). Penyakit diabetes sering menimbulkan komplikasi berupa stroke, gagal ginjal, jantung, nefropati, kebutaan dan bahkan harus menjalani amputasi jika anggota badan menderita luka gangrene (Algoritma *et al*., 2020). dimana kondisi luka ini biasanya terjadi di tungkai, jari kaki, dan jari tangan.

BAB III

METODOLOGI PENELITIAN

3.1 Tempat

Penelitian ini dilakukan di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam dan Perpustakaan Universitas Sriwijaya.

3.2 Waktu

Waktu penulisan penelitian ini dilakukan pada bulan Desember 2022 sampai dengan bulan Maret 2023.

3.3 Metode Penelitian

Langkah-langkah yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Mendeskripsikan data penyakit diabetes, data yang digunakan pada penelitian ini adalah data dari *dataset* kaggel. File data bernama diabetes.csv variabel yang digunakan pada penelitian ini adalah sebanyak 17 variabel dengan jumlah dataset sebanyak 520. Ini termasuk data tentang gejala penyebab diabetes.

   Link data: https://www.kaggle.com/*datasets*/yasserhessein/early-stage-diabetes-risk-prediction-*dataset*
2. Melakukan diskritisasi data penyakit diabetes. Proses diskritisasi dilakukan dengan konversi data numerik menjadi data kategori untuk variabel X_1 (Umur).

3. Melakukan partisi data menggunakan teknik *split validation* membagi data *training* menjadi 70% sebanyak 364 data dan data *testing* 30% sebanyak 156 data.
4. a. Melakukan perhitungan dengan metode *Naïve Bayes* pada klasifikasi penyakit diabetes:
    1. Menghitung P(Y) yaitu peluang variabel dependen (diabetes) dari setiap kategori yang disebut *perior*.
    2. Menghitung $P(X_1, \dots, X_n|Y)$ yaitu peluang dengan kategori variabel independen (umur, jenis kelamin, *polyuria*, *polydipsia*, berat badan turun, keletihan, *polyphagia*, iritasi genital, penglihatan kabur, gatal, mudah marah, sembuh lambat, *partial paresis*, otot kaku, rambut rontok, obesitas) terhadap variabel dependen (diabetes) yang disebut *likelihood*.
    3. Menghitung $P(Y|X_1, \dots, X_n)$ atau posterior dengan cara mengalikan semua peluang *prior* dan *likelihood* untuk masing-masing kategori variabel independen.
    4. Melakukan klasifikasi penyakit diabetes berdasarkan peluang *posterior* terbesar dari setia kategori variabel dependen.
    5. Menghitung tingkat akurasi penyakit diabetes menggunakan *Confusion Matrix* pada data *testing*.

    b. Melakukan perhitungan dengan menggunakan algoritma *Iterative Dichotomiser Three (ID3).*

1. Menghitung nilai entropy dan nilai gain menggunkan persamaan (2.3) dan (2.4)
2. Nilai gain tertinggi digunakan sebagai *root node*.
3. Selanjutnya ulangi tahapan diatas untuk membentuk pohon keputusan.
4. Melakukan klasifikasi dengan pohon keputusan yang telah terbentuk.
5. Menghitung tingkat akurasi menggunakan *Confusion Matrix* pada data *testing.*

5. Analisis hasil
6. Kesimpulan

BAB IV

HASIL DAN PEMBAHASAN

4.1 Deskripsi Data

Pada penelitian ini menggunakan data sekunder yang diambil dari *Kaggle.com* mengenai penyakit diabetes. File data bernama diabetes.csv Variabel yang digunakan pada penelitian ini adalah sebanyak 16 variabel independen dan 1 variabel dependen dengan jumlah dataset sebanyak 520. Berikut ini deskripsi variabel penelitian dapat dilihat pada tabel 4.1.

Tabel 4.1 Deskripsi variabel independen dan dependen pada data penyakit diabetes.

No	Variabel	Skala	Data
1	X_1 (Umur)	Rasio	16-90
2	X_2 (Jenis Kelamin)	Normal	1. Laki-Laki
			2. Perempuan
3	X_3 (*Polyuria*)	Normal	1. Ya
			0. Tidak
4	X_4 (*Polydispia*)	Normal	1. Ya
			0. Tidak
5	X_5 (Berat Badan Turun)	Normal	1. Ya
			0. Tidak
6	X_6 (Keletihan)	Normal	1. Ya
			0. Tidak
7	X_7 (*Polyphagia*)	Normal	1. Tidak
			0. Ya
8	X_8 (Iritasi Genetik)	Normal	1. Ya
			0. Tidak
9	X_9 (Penglihatan Kabur)	Normal	1. Ya
			0. Tidak
10	X_{10} (Gatal)	Normal	1. Ya
			0. Tidak

No	Variabel	Skala	Data
11	X_{11} (Mudah Marah)	Normal	1. Ya
			0. Tidak
12	X_{12} (Sembuh Lambat)	Normal	1. Ya
			0. Tidak
13	X_{13} (*Partial Paresis*)	Normal	1. Ya
			0. Tidak
14	X_{14} (Otot Kaku)	Normal	1. Ya
			0. Tidak
15	X_{15} (Rambut Rontok)	Normal	1. Ya
			0. Tidak
16	X_{16} (Obesitas)	Normal	1. Ya
			0. Tidak
17	Y (Diabetes)	Normal	1. Positif
			0. Negatif

4.2 Diskritisasi Data

Diskritisasi data dapat digunkan untuk konversi variabel kontinu menjadi variabel kategorik. Pada dataset yang digunakan terdapat satu variabel yang akan di diskritisasikan yaitu variablel X_1(umur). Diskritisasi pada data diabetes dapat dilihat pada tabel 4.2.

Tabel 4.2 Diskritisasi data pada variabel independen

Variabel	Kategori	Label	Interval
X_1(umur)	1	Remaja	11-19
	2	Dewasa	20-59
	3	Lansia	>60

(Makmun & Radisu, 2021).

4.3 Partisi Data

Partisi data dapat digunakan untuk membagi dataset ke dalam data *training* dan data *testing*. Teknik yang digunakan yaitu menggunakan teknik *split validation*

dimana data di partisi ke dalam 70% data *training* atau sebanyak 364 data dan 30% data *testing* atau 156 data.

Table 4.3 Data *training* 70% pada penyakit diabetes

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	Y
1	2	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
2	2	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	2	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	3	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
5	2	1	1	1	0	1	1	1	0	0	0	1	1	0	0	0	1
⋮																	
364	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4.4 Data *testing* 30% pada penyakit diabetes

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	Y
1	2	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	2	1	1	1	0	1	1	0	1	1	0	1	0	1	1	1	1
3	3	1	0	1	1	1	1	0	1	1	1	0	0	0	1	0	1
4	2	1	1	1	0	0	1	1	0	1	0	1	0	1	0	0	1
5	3	1	1	1	0	1	1	0	1	0	0	0	1	1	0	0	1
⋮																	
156	3	2	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1

4.4 Metode Naïve Bayes

Pada 520 dataset terdapat 200 orang negatif diabetes dan 320 orang positif diabetes. Proses perhitungan metode Naïve Bayes dilakukan dengan menggunakan data *training*. Dari 364 data *training* terdapat 140 orang negatif diabetes dan 224 positif diabetes, sehingga dapat dicari peluang *prior* dari data *training* adalah sebagai berikut.

$$P\left(Y_{negatif} = 0\right) = \frac{n(Y=0)}{n_{total}} = \frac{140}{364} = 0{,}384615$$

$$P\left(Y_{positif} = 1\right) = \frac{n(Y=2)}{n_{total}} = \frac{244}{364} = 0{,}615385$$

Selanjutnya, menghitung peluang untuk setiap kategori dari variabel independen terhadap variabel dependen *(likelihood)* sehingga diperoleh hasil sebagai berikut :

$$P(X|Y_i) = \frac{P(Y_i|X)}{P(X)}$$

$$P\left(X_{umur=1}\middle|D_{negatif}\right) = \frac{0}{140} = 0$$

$$P\left(X_{umur=1}\middle|D_{positif}\right) = \frac{0}{224} = 0$$

$$P\left(X_{umur=2}\middle|D_{negatif}\right) = \frac{120}{140} = 0{,}8571$$

$$P\left(X_{umur=2}\middle|D_{positif}\right) = \frac{173}{224} = 0{,}7723$$

$$P\left(X_{umur=3}\middle|D_{negatif}\right) = \frac{20}{140} = 0{,}1428$$

$$P\left(X_{umur=3}\middle|D_{positif}\right) = \frac{51}{224} = 0{,}2276$$

Penyelesaian diatas menggunakan dua sampel data dengan variabel independen yaitu umur dengan kelas pertama, kemudian dengan cara yang sama menghitung setiap variabel untuk masing-masing kelasnya, perhitungan dapat dilihat pada tabel sebagai berikut.

Tabel 4.5 perhitungan nilai *likelihood*

		jumlah kejadian Y		probabilitas	
variabel	kategorik	0 (negatif)	1(positif)	0 (negatif)	1 (positif)
X_1	1	0	0	0	0
	2	120	173	0,857143	0,772321
	3	20	51	0,142857	0,227679
Jumlah		140	224	1	1
X_2	1	128	104	0,914286	0,464286
	2	12	120	0,085714	0,535714
Jumlah		140	224	1	1
⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	118	180	0,842857	0,803571
	1	22	44	0,157143	0,196429
Jumlah		140	224	1	1

Setelah didapat hasil *likelihood* dengan menggunakan *data training*, selanjutnya mengklasifikasikan penyakit diabetes dengan cara mengalikan semua peluang *prior* dan *likelihood* dari masing-masing kategorik variabel independen.

Berikut merupakan perhitungan posterior dari data *testing* sebagai berikut:

$$P(Y=0|X_1,X_2,\dots X_{16}) = P(X_1=2|Y=0)P(X_2=1|Y=0)P(X_3=0|Y=0)$$
$$P(X_4=0|Y=0)P(X_5=0|Y=0)P(X_6=1|Y=0)$$
$$P(X_7=0|Y=0)P(X_8=0|Y=0)P(X_9=1|Y=0)$$
$$P(X_{10}=0|Y=0)P(X_{11}=0|Y=0)P(X_{12}=0|Y=0)$$
$$P(X_{13}=1|Y=0)P(X_{14}=0|Y=0)P(X_{15}=1|Y=0)$$
$$P(X_{16}=0|Y=0)P(Y=0)$$
$$= (0{,}8571)(0{,}9143)(0{,}9143)(0{,}9643)(0{,}8357)$$
$$(0{,}4286)(0{,}7929)(0{,}8214)(0{,}2500)(0{,}5357)$$
$$(0{,}9071)(0{,}5929)(0{,}1643)(\,0{,}7214)(0{,}4714)$$
$$(0{,}8429)(0{,}3846)$$
$$= \ 0{,}000210$$

$$P(Y=1|X_1,X_2,\dots X_{16}) = P(X_1=2|Y=1)P(X_2=1|Y=1)P(X_3=0|Y=1)$$
$$P(X_4=0|Y=1)P(X_5=0|Y=1)P(X_6=1|Y=1)$$
$$P(X_7=0|Y=1)P(X_8=0|Y=1)P(X_9=1|Y=1)$$
$$P(X_{10}=0|Y=1)P(X_{11}=0|Y=1)P(X_{12}=0|Y=1)$$
$$P(X_{13}=1|Y=1)P(X_{14}=0|Y=1)P(X_{15}=1|Y=1)$$
$$P(X_{16}=0|Y=1)P(Y=1)$$
$$= (0{,}7723)(0{,}4643)(0{,}2366)(0{,}2857)(0{,}4018)$$
$$= (0{,}6696)(0{,}3929)(0{,}7366)(0{,}5268)(0{,}4911)$$
$$(0{,}6339)(0{,}5089)(0{,}5982)(0{,}5669)(0{,}25446)$$

$$(0{,}8036)(0{,}6154)$$

$$= 0{,}00000672261$$

Berdasarkan hasil perhitungan diatas, maka $P(Y = 0|X_1, X_2, \ldots X_{16})$ merupakan nilai peluang tertinggi untuk data *testing* pertama yang terklasifikasi pada kategori pertama yaitu kelompok negatif. Kemudian selanjutnya mencari nilai posterior seperti perhitungan diatas untuk data *testing* selanjutnya sampai terakhir serta klasifikasikan berdasarkan nilai posterior tertinggi. Hasil lengkap perhitungan nilai *posterior* dapat dilihat pada tabel 4.6.

Tabel 4.6 perhitungan nilai *posterior*

No	Perhitungan nilai *posterior*	
	$P(Y = 0\|X_1, X_2, \ldots X_{16})$	$P(Y = 1\|X_1, X_2, \ldots X_{16})$
1	0,00021026	6,72261E-06
2	4,15748E-08	1,05085E-05
3	1,80493E-07	1,55443E-05
4	2,1741E-07	1,99588E-05
6	8,25792E-08	0,000187527
7	5,33846E-07	5,10307E-05
8	2,0785E-09	6,58558E-05
9	7,84521E-07	3,95419E-05
10	1,08267E-08	0,000180367
11	5,03106E-09	3,87273E-05
12	1,31897E-09	0,000333877
13	1,08011E-07	8,04573E-05
⋮	⋮	⋮
156	2,81259E-09	6,6716E-05

4.5 *Confusion Matrix* metode *Naïve Bayes*

Untuk mengukur kinerja klasifikasi penyakit diabetes dapat dilakukan dengan menggunakan *confusion matrix* berdasarkan nilai akurasi yang dapat dilihat pada tabel 4.7 sebagai berikut:

Tabel 4.7 *Confusion Matrix* metode *Naïve Bayes*

Prediksi	Class	
	Positif (1)	Negatif (0)
Positif (1)	83	5
Negatif (0)	13	55

Berdasarkan tabel diatas, dapat dilihat bahwa terdapat 83 orang yang diprediksi benar positif diabetes dan 55 orang diprediksi benar negatif diabetes. Kemudian 5 orang diprediksi positif diabetes tetapi pada data sebenarnya negatif diabetes dan 13 orang diprediksi negatif diabetes tetapi pada data sebenarnya positif diabetes. Selanjutnya dapat dihitung tingkat akurasi pada tabel diatas sebagai berikut:

$$akurasi = \frac{83 + 55}{83 + 5 + 13 + 55} 100\% = 0{,}8846 = 88{,}46\%$$

$$recall \quad = \frac{83}{83 + 13} 100\% = 0{,}8646 = 86{,}46\%$$

$$precision = \frac{83}{83 + 5}\ 100\% = 0{,}9432 = 94{,}32\%$$

4.6 *Algoritma Iterative Dichotomiser Three (ID3)*

Proses perhitungan algoritma *Iterative Dichotomiser Three (ID3)* dilakukan dengan menggunakan data *training*. Dari 364 data *training* terdapat 140 orang negatif diabetes dan 224 positif diabetes, sehingga dapat dicari peluang dari data *training* adalah sebagai berikut.

$$P\left(Y_{negatif} = 0\right) = \frac{n(Y=0)}{n_{total}} = \frac{140}{364} = 0{,}384615$$

$$P\left(Y_{positif} = 1\right) = \frac{n(Y=2)}{n_{total}} = \frac{244}{364} = 0{,}615385$$

Mencari nilai entropy untuk masing-masing nilai atribut sebagai berikut:

Dari keseluruhan data *training* terdapat 244 kelas untuk diabetes = positif dan 140 instance untuk diabetes = negatif. Berikut merupakan perhitunganya:

$$Entropy(S) = -\sum_i p_i \, log_2 \, p_i$$

$$Entropy(S) = \left(-\left(\frac{140}{364}\right) log_2 \left(\frac{140}{364}\right)\right) + \left(-\left(\frac{244}{364}\right) log_2 \left(\frac{244}{364}\right)\right)$$

$$Entropy = 0{,}9612$$

$$Entropy \;\; (X_1 = 1) = \left(-\left(\frac{0}{0}\right) log_2 \left(\frac{0}{0}\right)\right) + \left(-\left(\frac{0}{0}\right) log_2 \left(\frac{0}{0}\right)\right)$$

$$Entropy = 0 \; (tidak \; terdefinisi)$$

$$Entropy \; (X_1 = 2) = \left(-\left(\frac{120}{293}\right) log_2 \left(\frac{120}{293}\right)\right) + \left(-\left(\frac{173}{293}\right) log_2 \left(\frac{173}{293}\right)\right)$$

$$Entropy = 0{,}9763$$

$$Entropy \; (X_1 = 3) = \left(-\left(\frac{20}{71}\right) log_2 \left(\frac{20}{71}\right)\right) + \left(-\left(\frac{51}{71}\right) log_2 \left(\frac{51}{71}\right)\right)$$

$$Entropy = 0{,}8577$$

$$\vdots$$

$$Entropy \; (X_{16=0}) = \left(-\left(\frac{118}{298}\right) log_2 \left(\frac{118}{298}\right)\right) + \left(-\left(\frac{180}{298}\right) log_2 \left(\frac{180}{298}\right)\right)$$

$$Entropy = 0{,}9685$$

$$Entropy \; (X_{16=1}) = \left(-\left(\frac{22}{66}\right) log_2 \left(\frac{22}{66}\right)\right) + \left(-\left(\frac{44}{66}\right) log_2 \left(\frac{44}{66}\right)\right)$$

$$Entropy = 0{,}9182$$

Setelah mencari nilai entropy diatas dari setiap variabel langkah selanjutnya menghitung nilai information gain untuk semua variabel independen agar mendapatkan nilai gain tertinggi.

$$nformation\ gain\ (S,A) = entropy(S) - \sum_{i} \frac{S_i}{S} x\ entropy\ (S_i)$$

$$Gain(S, x_1) = Entropy(S) - \sum_{i=1}^{3} \frac{x_{1i}}{n_{total}} Entropy(x_{1i})$$

$$= 0{,}9612 - \left(\left(\frac{293}{364}\right)0{,}9763 + \left(\frac{71}{364}\right)0{,}8577\right)$$

$$= 0{,}0081$$

$$Gain(S, x_2) = Entropy(S) - \sum_{i=1}^{2} \frac{x_{2i}}{n_{total}} Entropy\ (x_{2i})$$

$$= 0{,}9612 - \left(\left(\frac{232}{364}\right)0{,}9923 + \left(\frac{132}{364}\right)0{,}4395\right)$$

$$= 0{,}1694$$

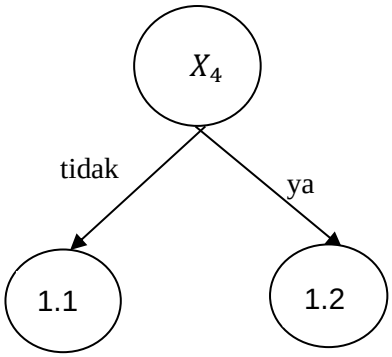
$$\vdots$$

$$Gain(S, x_{16}) = Entropy(S) - \sum_{i=1}^{2} \frac{x_{16i}}{n_{total}} Entropy(x_{16i})$$

$$= 0{,}9612 - \left(\left(\frac{298}{364}\right)0{,}9685 + \left(\frac{66}{364}\right)0{,}9183\right)$$

$$= 0{,}0018$$

Setelah dilakukan perhitungan gain dari setiap nilai entropy dapat dilihat dari hasil perhitungan diatas, nilai gain tertinggi terdapat pada $X_4(Polydispia)$ sebesar 0,3771 maka X_4 $(Polydispia)$ adalah node pertama atau *rote node*. Selanjutnya membuat cabang dari X_4 dimana memiliki 2 kategori, kategori tersebut terdiri dari (ya) dan (tidak). Kemudian dapat membuat gambar pohon keputusan *rote node* pada gambar 4.1.



Gambar 4.1 pohon keputusan *root node*

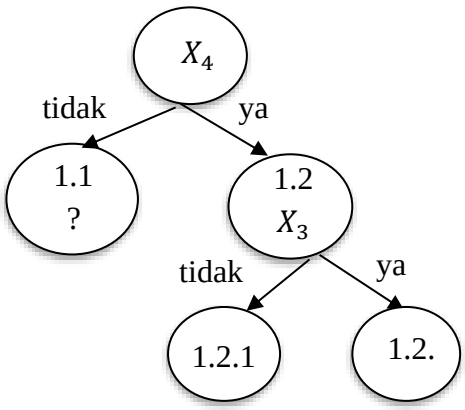
Langkah selanjutnya melakukan perhitungan pada semua cabang X_4 dalam hal ini perhitungan dilakukan pada kategori (ya). Menghitung variabel X_4 dengan kategori (ya) pada seluruh data terdapat 165 kasus yang terdiri dari 5 orang negatif diabetes dan 160 orang positif diabetes. Kemudian hitung kembali nilai entropy dan gain dengan cara sebelumnya tetapi tanpa menghitung pada variabel X_4. Perhitungan entropy dan gain dapat dilihat pada tabel

Tabel 4.8 perhitungan entropy dan gain variabel X_4 kategori ya (1.2)

variabel	Kategori	Jumlah	0 (negatif)	1(positif)	Entropy	Gain
total		165	5	160	0,1959	
X_1	1	0	0	0	0	
	2	129	5	124	0,2366	0,0109
	3	36	0	36	0	
X_2	1	78	5	73	0,3435	0,0335
	2	87	0	87	0	
X_3	0	28	5	23	0,6769	0,0810
	1	137	0	137	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	129	4	125	0,1994	4,46e-05
	1	36	1	35	0,1831	

Berdasarkan tabel diatas terlihat bahwa nilai gain tertinggi terdapat pada variabel X_3 (Polyuria) sebesar 0,0810 sehingga X_3 merupan percabangan dari variabel X_4 yang disebut dengan note 1.2. pada variabel X_3 memiliki 2 kategori yang mana ini berarti

X_3 memiliki 2 cabang. Pohon keputusan pada node 1.2 dapat dapat dilihat pada gambar berikut.



Gambar 4.2 pohon keputusan untuk note 1.2

Selanjutnya menghitung kembali untuk cabang X_3 pada kategori (ya) untuk mendapatkan note berikutnya. Menghitung variabel X_3 dengan kategori (ya) pada seluruh data (137), terdapat 137 kasus dimana tidak terdapat orang negatif diabetes namun terdapat 137 orang positif diabetes. Kemudian menghitung variabel X_3 dengan kategorik (tidak) pada seluruh data (28) dimana terdapat 5 orang negatif diabetes dan 23 orang positif diabetes langkah selanjutnya hitung kembali nilai entropy dan informasi gain dengan cara sebelumnya tetapi tanpa menghitung variabel X_4 dan X_3 Perhitungan entropy dan information gain dapat dilihat pada tabel berikut:

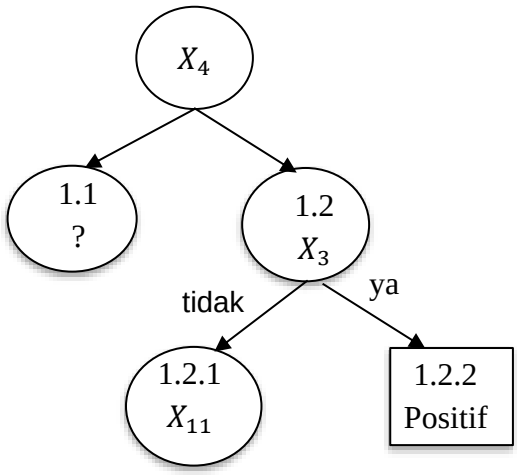
Tabel 4.9 perhitungan entropy dan gain variabel X_3 kategori tidak (1.2.1)

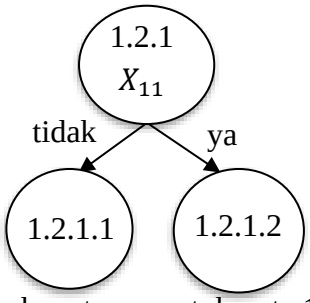
Variabel	Kategori	Jumlah	0 (negatif)	1(positif)	entropy	Gain
Total		28	5	23	0,6769	
X_1	1	0	0	0	0	
	2	21	5	16	0,7919	0,0830
	3	7	0	7	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	21	4	17	0,7025	
	1	7	1	6	0,5917	0,0021

Tabel 4.10 perhitungan entropy dan gain variabel X_3 kategori ya (1.2.2)

Variabel	Kategori	jumlah	0 (negatif)	1(positif)	entropy	gain
		137	0	137	0	
X_1	1	0	0	0	0	0
	2	108	0	108	0	
	3	29	0	29	0	
X_2	1	59	0	59	0	0
	2	78	0	78	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	108	0	108	0	0
	1	29	0	29	0	

Berdasarkan tabel 4.9 diatas terlihat bahwa nilai gain tertinggi terdapat pada variabel X_{11} (Mudah Marah) sebesar 0,01850 sehingga X_{11} merupakan percabangan dari variabel X_3 kategori tidak yang disebut dengan note 1.2.1 pada variabel X_{11} memiliki 2 kategori yang mana ini berarti X_{11} memiliki 2 cabang. Sedangkan pada tabel 4.10 terlihat bahwa semua nilai gain bernilai 0 ini berarti untuk gejala pada variabel X_3 (Polyuria) dengan kategori (ya) tidak dapat dilakukan percabangan kembali yang mana ini berarti seseorang positif diabetes. Pohon keputusan pada node 1.2.1 dan node 1.2.2 dapat dapat dilihat pada gambar berikut.





Gambar 4.3 pohon keputusan untuk note 1.2.1 dan 1.2.2

Selanjutnya menghitung kembali untuk cabang X_{11} pada kategori (tidak) untuk mendapatkan note berikutnya. Perhitungan dilakukan dengan memfilter data, pada variabel X_{11} dengan kategori tidak perhitungan entropy dan information gain dapat dilihat pada tabel berikut:

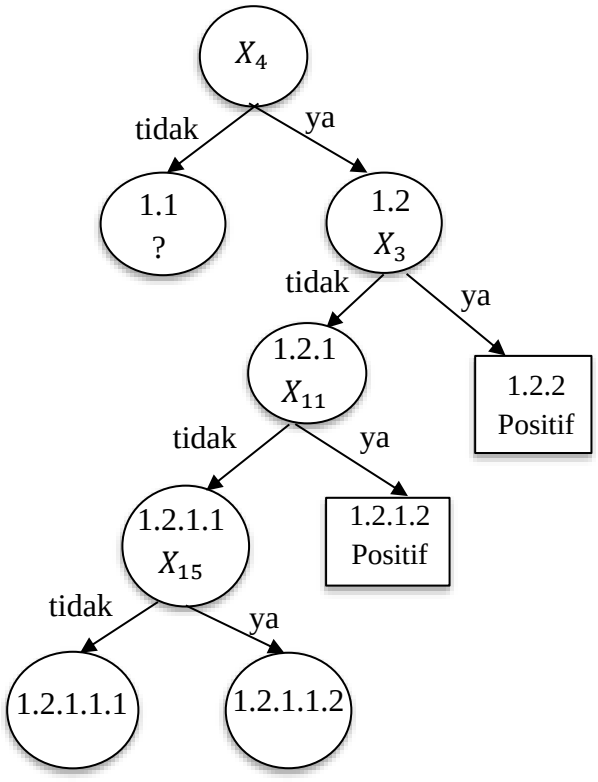
Tabel 4.11 perhitungan entropy dan gain variabel X_{11} kategori tidak (1.2.1.1)

variabel	kategori	jumlah	0 (negatif)	1(positif)	entropy	gain
Total		15	5	10	0,9183	
X_1	1	0	0	0	0	
	2	13	5	8	0,9612	0,0852
	3	2	0	2	0	
X_2	1	11	5	6	0,994	0,1893
	2	4	0	4	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	11	4	7	0,9457	
	1	4	1	3	0,8113	0,0085

Tabel 4.12 perhitungan entropy dan gain variabel X_{11} kategori ya (1.2.1.2)

Variabel	Kategori	jumlah	0 (negatif)	1(positif)	entropy	gain
Total		13	0	13	0	
X_1	1	0	0	0	0	0
	2	8	0	8	0	
	3	5	0	5	0	
X_2	1	8	0	8	0	0
	2	5	0	5	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	10	0	10	0	0
	1	3	0	3	0	

Berdasarkan tabel 4.11 diatas terlihat bahwa nilai gain tertinggi terdapat pada variabel X_{15} (rambut rontok) sebesar 0,4093 sehingga X_{15} merupakan percabangan dari variabel X_{11} dengan kategori (tidak) yang disebut dengan note 1.2.1.1 pada variabel X_{15} memiliki 2 kategorik yang mana ini berarti X_{15} memiliki 2 cabang. Sedangkan pada tabel 2.12 terlihat bahwa semua nilai gain bernilai 0 ini berarti untuk gejala pada variabel X_{11} (rambut rontok) dengan kategori (ya) tidak dapat dilakukan percabangan kembali yang mana ini berarti seseorang positif diabetes. Pohon keputusan pada node 1.2.1.1 dan node 1.2.1.2 dapat dapat dilihat pada gambar berikut:



Gambar 4.4 pohon keputusan untuk note 1.2.1.1 dan 1.2.1.2

Selanjutnya menghitung kembali untuk cabang X_{15} pada kategori (tidak) untuk mendapatkan note berikutnya. Perhitungan dilakukan dengan memfilter data, pada variabel X_{15} dengan kategori (tidak) Perhitungan entropy dan information gain dapat dilihat pada tabel berikut:

Tabel 4.13 perhitungan entropy dan gain variabel X_{15} kategori tidak (1.2.1.1. 1)

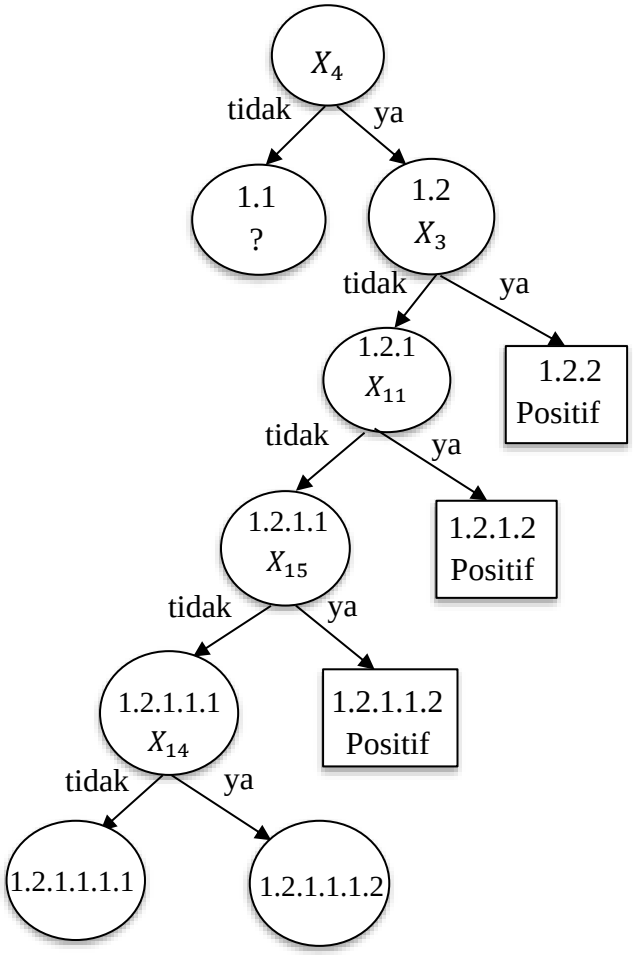
Variabel	Kategori	jumlah	0 (negatif)	1(positif)	entropy	gain
total		8	5	3	0,954	
X_1	1	0	0	0	0	
	2	8	5	3	0,954	0
	3	0	0	0	0	
X_2	1	6	5	1	0,65	0,4669
	2	2	0	2	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	7	4	3	0,985	0,0924
	1	1	1	0	0	

Tabel 4.14 perhitungan entropy dan gain variabel X_{15} kategori ya (1.2.1.1. 2)

Variabel	Kategori	jumlah	0 (negatif)	1(positif)	entropy	gain
Total		7	0	7	0	
X_1	1	0	0	0	0	0
	2	5	0	5	0	
	3	2	0	2	0	
X_2	1	5	0	5	0	0
	2	2	0	2	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	4	0	4	0	0
	1	3	0	3	0	

Berdasarkan tabel 4.13 terlihat bahwa nilai gain tertinggi terdapat pada variabel X_{14} (otot kaku) sebesar 0,5488 sehingga X_{14} merupakan percabangan dari variabel X_{15} pada kategori (tidak) yang disebut dengan note 1.2.1.1.1 pada variabel X_{14}

memiliki 2 kategori yang mana ini berarti X_{14} memiliki 2 cabang. Sedangkan pada tabel 4.14 terlihat bahwa semua nilai gain bernilai 0 ini berarti untuk gejala pada variabel X_{15} (rambut rontok) dengan kategori (ya) tidak dapat dilakukan percabangan kembali yang mana ini berarti seseorang positif diabetes. Pohon keputusan pada node 1.2.1.1.1 dan node 1.2.1.1.2 dapat dilihat pada gambar berikut:



Gambar 4.5 pohon keputusan untuk note 1.2.1.1.1 dan 1.2.1.1.2

Selanjutnya menghitung kembali untuk cabang X_{14} pada kategori (tidak) untuk mendapatkan note berikutnya. Perhitungan dilakukan dengan memfilter data, pada variabel X_{14} dengan kategori tidak. Perhitungan entropy dan information gain dapat dilihat pada tabel berikut.

Tabel 4.15 perhitungan entropy dan gain variabel X_{14} kategori tidak (1.2.1.1.1.1)

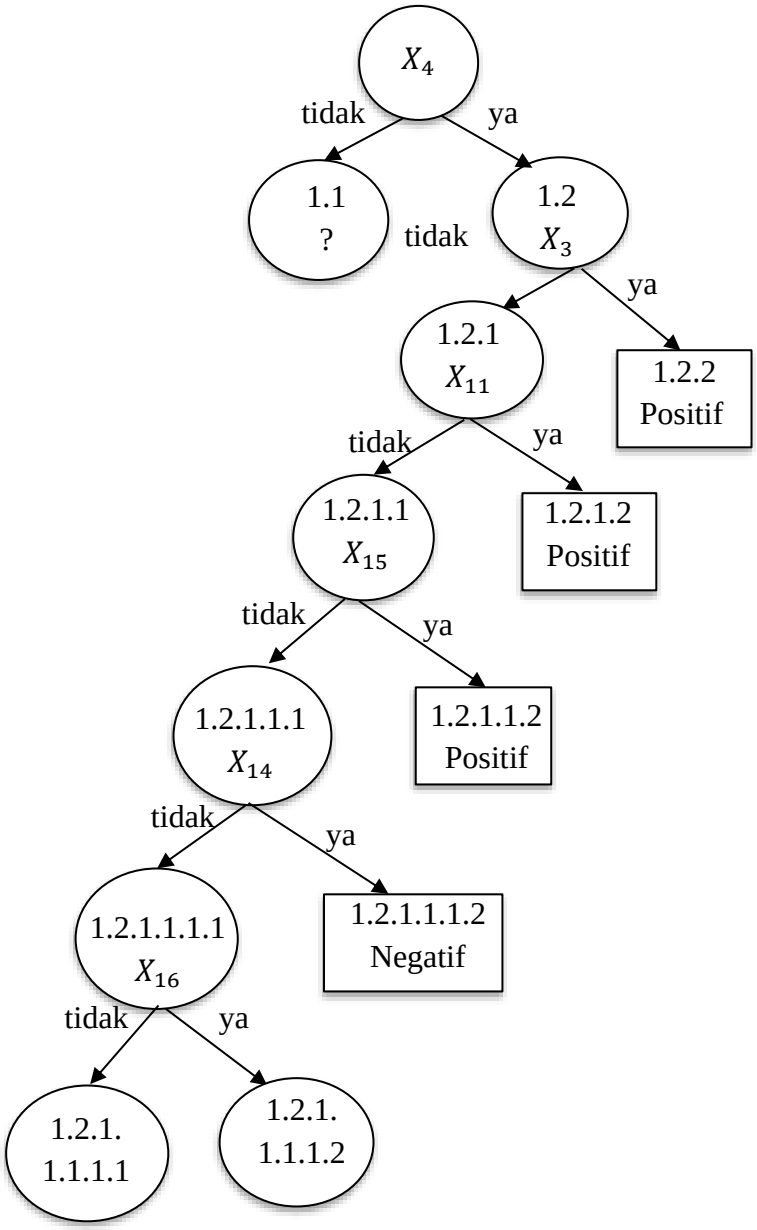
Variabel	Kategori	Jumlah	0 (negatif)	1(positif)	entropy	gain
Total		4	1	3	0,8113	
X_1	1	0	0	0	0	0
	2	4	1	3	0,8113	
	3	0	0	0	0	
X_2	1	2	1	1	1	0,3113
	2	2	0	2	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	3	0	3	0	0,8113
	1	1	1	0	0	

Tabel 4.16 perhitungan entropy dan gain variabel X_{14} kategori ya (1.2.1.1.1.2)

Variabel	Kategori	Jumlah	0 (negatif)	1(positif)	entropy	gain
Total		4	4	0	0	
X_1	1	0	0	0	0	0
	2	4	4	0	0	
	3	0	0	0	0	
X_2	1	4	4	0	0	0
	2	0	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{16}	0	4	4	0	0	0
	1	0	0	0	0	

Berdasarkan tabel terlihat bahwa nilai gain tertinggi terdapat pada variabel X_{16} (obesitas) sebesar 0,8113 sehingga X_{16} merupakan percabangan dari variabel X_{14} pada kategori (tidak) yang disebut dengan note 1.2.1.1.1.1 pada variabel X_{16} memiliki 2 kategori yang mana ini berarti X_{16} memiliki 2 cabang. Sedangkan pada tabel 4.16 terlihat bahwa semua nilai gain bernilai 0 ini berarti untuk gejala pada variabel X_{16} (obesitas) dengan kategori (ya) tidak dapat dilakukan percabangan kembali yang mana ini berarti seseorang negatif diabetes ini terlihat pada jumlah

peluang negatif pada tabel. Pohon keputusan pada node 1.2.1.1.1.1 dan node 1.2.1.1.1.2 dapat dapat dilihat pada gambar berikut:



Gambar 4.6 pohon keputusan untuk note 1.2.1.1.1.1 dan 1.2.1.1.1.2

Selanjutnya menghitung kembali untuk cabang X_{16} pada kategori (tidak) dan kategori untuk mendapatkan note berikutnya. Perhitungan dilakukan dengan

memfilter data, pada variabel X_{16} dengan kategori (tidak) dan (ya). Perhitungan entropy dan information gain dapat dilihat pada tabel berikut:

Tabel 4.17 perhitungan entropy dan gain variabel X_{16} kategori tidak (1.2.1.1.1.1.1)

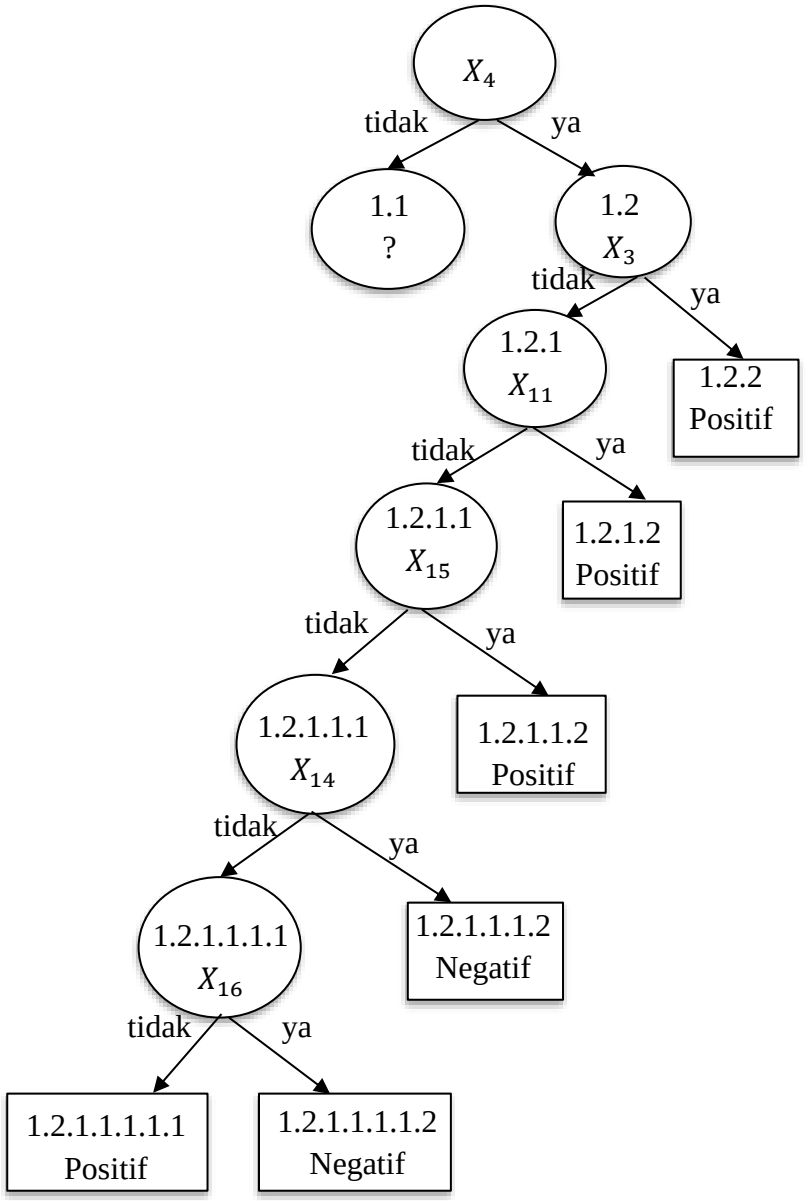
Variabel	Kategori	jumlah	0 (negatif)	1(positif)	entropy	gain
Total		3	0	3	0	
X_1	1	0	0	0	0	0
	2	3	0	3	0	
	3	0	0	0	0	
X_2	1	1	0	1	0	0
	2	2	0	2	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{13}	0	2	0	2	0	0
	1	0	0	1	0	

Tabel 4.18 perhitungan entropy dan gain variabel X_{16} kategori ya (1.2.1.1.1.1.2)

Variabel	Kategori	jumlah	0 (negatif)	1(positif)	entropy	gain
Total		1	1	0	0	
X_1	1	0	0	0	0	0
	2	1	1	0	0	
	3	0	0	0	0	
X_2	1	1	1	0	0	0
	2	0	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{13}	0	0	0	0	0	0
	1	1	1	0	0	

Berdasarkan tabel 4.17 dan tabel 4.18 terlihat bahwa semua gain bernilai 0 ini berarti pada variabel X_{16} (obesitas) untuk kategorik (tidak) dan (ya) tidak dapat lagi melakukan percabangan ini berarti telah mencapai titik akhir yang disebut node daun (*leaf node*) pada kategorik (tidak) untuk gejala variabel X_{16} (obesitas) seseorang positif diabetes terlihat dari jumlah seseorang positif diabetes pada tabel 4.17 Kemudian untuk variabel X_{16} (obesitas) dengan kategorik (ya) seseorang

dikatakan negatif diabetes ini terlihat pada jumlah seseorang negatif pada tabel 4.18. Pohon keputusan pada node 1.2.1.1.1.1.1 dan node 1.2.1.1.1.1.2 dapat dilihat pada gambar berikut:



Gambar 4.7 pohon keputusan untuk note 1.2.1.1.1.1.1 dan 1.2.1.1.1.1.2

Berdasarkan gambar 4.7 diperoleh model dengan menggunakan data *training* yang mana model tersebut memiliki aturan klasifikasi sebagai berikut:

1. Jika X_4(polydipsia) = ya dan X_3 (polyuria) = ya maka prediksinya positif diabetes.
2. Jika X_4(polydipsia) = ya dan X_3(polyuria) = tidak dan X_{11}(mudah marah) = ya maka prediksinya positif diabetes.
3. Jika X_4(polydipsia) = ya dan X_3(polyuria) = tidak dan X_{11}(mudah marah) = tidak dan X_{15}(rambut rontok) = ya maka prediksinya positif diabetes.
4. Jika X_4(polydipsia) = ya dan X_3(polyuria) = tidak dan X_{11}(mudah marah) = tidak dan X_{15}(rambut rontok) = tidak dan X_{14}(otot kaku) = ya maka prediksinya negatif diabetes.
5. Jika X_4(polydipsia) = ya dan X_3(polyuria) = tidak dan X_{11}(mudah marah) tidak dan X_{15}(rambut rontok) = tidak dan X_{14}(otot kaku) = tidak dan X_{16} (obesitas) = ya maka prediksinya negatif diabetes
6. Jika X_4(polydipsia) = ya dan X_3(polyuria) = tidak dan X_{11}(mudah marah) tidak dan X_{15}(rambut rontok) = tidak dan X_{14}(otot kaku) = tidak dan X_{16}(obesitas) = tidak maka prediksinya positif diabetes

Proses pohon keputusan diatas dilakukan juga terhadap variabel X_4 (polydipsia) dengan kategori (tidak) untuk mendapatkan percabangan 1.1 ulangi proses diatas sampai tidak ada lagi cabang yang tersisa atau semua cabang sudah mencapai titik akhir *(leaft node*). Jika satu pohon keputusan sudah didapat, selanjutnya membuat pengkondisian pohon, dimana kondisi ini digunakan sebagai prediksi klasifikasi terhadap data *testing*. Pohon keputusan dapat dilihat pada lampiran .

4.7 *Confusion Matrix* Algoritma *Iterative Dichotomiser Three (ID3)*

Untuk mengukur kinerja klasifikasi dengan algoritma *Iterative Dichotomiser Three (ID3)* dapat dilakukan dengan menggunakan *confusion matrix* berdasarkan nilai akurasi yang dapat dilihat pada tabel 4.19 sebagai berikut:

Tabel 4.19 *Confusion Matrix* algoritma *Iterative Dichotomiser Three (ID3)*

Prediksi	Class	
	Positif (1)	Negatif (0)
Positif (1)	92	3
Negatif (0)	4	57

Berdasarkan tabel diatas, dapat dilihat bahwa terdapat 92 orang yang diprediksi benar positif diabetes dan 57 orang diprediksi benar negatif diabetes. Kemudian 3 orang diprediksi positif diabetes tetapi pada data sebenarnya negatif diabetes dan 13 orang diprediksi negatif diabetes tetapi pada data sebenarnya positif diabetes. Selanjutnya dapat dihitung tingkat akurasi pada tabel diatas sebagai berikut:

$$akurasi = \frac{92 + 57}{92 + 3 + 4 + 57} 100\% = 0{,}9551 = 95{,}51\%$$

$$recall = \frac{92}{92 + 4} 100\% = 0{,}9583 = 95{,}83\%$$

$$precision = \frac{92}{92 + 3} 100\% = 0{,}9684 = 96{,}84\%$$

Hasil dari tingkat akurasi menggunakan algoritma *Iterative Dichotomiser Three (ID3)* dalam klasifikasi penyakit diabetes adalah 0,9551 atau 95,51%. Ini menunjukan bahwa ketepatan dalam memprediksi penyakit diabetes menggunakan algoritma *Iterative Dichotomiser Three (ID3)* sebesar 95,51%.

4.8 Analisis Hasil

Berdasarkan tingkat akurasi yang di dapat, metode *Naïve Bayes* memiliki tingkat akurasi dalam klasifikasi penyakit diabetes yaitu sebesar 88,46% recall 86,46% dan precision 94,32%. Proses klasifikasi metode *Naïve Bayes* diperoleh berdasarkan nilai *posterior* terbesar untuk menjadi penentu hasil klasifikasi positif atau negatif seseorang terkena diabetes. Sedangkan pada algoritma *Iterative Dichotomiser Three (ID3)* hasil klasifikasi dilihat berdasarkan pohon keputusan dengan cara mencari nilai *information* gain tertinggi sehingga diperoleh variabel X_4 (polydipsia) sebagai *root node* ini berarti X_4 (polydipsia) merupakan variabel yang paling berpengaruh pada proses klasifikasi penyakit diabetes. Hasil akurasi yang diperoleh pada algoritma *Iterative Dichotomiser Three (ID3)* yaitu sebesar 95,51% recall 95,83% dan precision 96,84%. Perbandingan dari metode *Naïve Bayes* dan algoritma *Iterative Dichotomiser Three (ID3)* dapat dilihat berdasarkan nilai akurasi. Algoritma *Iterative Dichotomiser Three* (*ID3)* memiliki tingkat akurasi yang lebih tinggi dibandingkan metode *Naïve Bayes.* Tingkat akurasi metode *Naïve Bayes* lebih rendah bisa terjadi karena *Naïve Bayes* memiliki asumsi bahwa masing-masing variabel independen bisa membuat berkurangnya akurasi. Maka algoritma *Iterative Dichotomiser Three* (*ID3)* dengan akurasi sebesar 95,51% lebih cocok digunakan pada data penelitian ini dari pada metode *Naïve Bayes*.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Proses klasifikasi penyakit diabetes menggunakan metode *Naïve Bayes* dihitung berdasarkan peluang dari setiap masing-masing variabel. Nilai terbesar dari peluang *posterior* akan menjadi penentu kelompok klasifikasi positif atau negatif pada penyakit diabetes, sedangkan pada algoritma *Iterative Dichotomiser Three (ID3)* proses klasifikasi dilihat berdasarkan pohon keputusan diperoleh X_4 (polydipsia) sebagai *root node*. X_4(polydipsia) merupakan variabel yang paling berpengaruh pada proses klasifikasi penyakit diabetes.
2. Hasil tingkat akurasi menggunakan metode *Naïve Bayes* yaitu sebesar 88,46% dengan recall 86,46% dan precision 94,32%. Sedangkan tingkat akurasi algoritma *Iterative Dichotomiser Three (ID3)* yaitu sebesar 95,51% dengan recall 95,83% dan precision 96,84% Ini menunjukan bahwa tingkat akurasi algoritma *Iterative Dichotomiser Three (ID3)* lebih cocok digunakan pada data penelitian ini.

5.2 Saran

Pada penelitian ini penulis menggunakan metode *Naïve Bayes* dan algoritma *Iterative Dichotomiser Three (ID3)*, sehingga untuk penelitian selanjutnya dapat menggunakan metode klasifikasi dan teknik partisi data yang dapat dilakukan dengan cara lainnya.

DAFTAR PUSTAKA.

Algoritma, C., Ente, D. R., Thamrin, S. A., & Kuswanto, H. (2020). Klasifikasi faktor-faktor penyebab penyakit diabetes melitus di rumah sakit unhas menggunakan algoritma *c4.5*. 80–88.

Budiyantara, A., Wijaya, A. K., Gunawan, A., & Rolland, M. (2021). Analisis data mining hotel booking menggunakan model Id3. *JBASE - Journal of Business and Audit Information Systems*, *4*(1), 1–12.

Laila, W. (2021). *Rekomendasi Makanan Bagi Pasien Hiperlipidiemia Berdasarkan Hasil Klasifikasi Menggunakan Metode Naive Bayes dan Decision Tree*. *8*(2), 328–336.

Makmun, A., & Radisu, I. M. (2021). Karakteristik pada Obesitas Berdasarkan Rentan Umur di Kelurahan Nganganaumala Kota Bau-Bau. *Indonesian Journal of Health*, *1*(2), 85–90.

Dwi, R., Prakoso, Y., & Sari, D. R. (2022). Perancangan Sistem Klasifikasi Penyakit Diabetes. 02(01), 24–31.

Ginting, R., Girsang, E., Bastira Ginting, J., studi kesehatan masyarakat Fakultas kedokteran, P., gigi, kedokteran, & ilmu kesehatan, dan. (2022). Analisis determinan dan prediksi penyakit diabetes melitus tipe 2 menggunakan metode machine learning: Scoping Review. *Jurnal.Unprimdn.Ac.Id*, *7*(1).

Giovani, R. A., Mudjihartono, P., & Pranowo, P. (2011). Sistem pendukung keputusan prediksi kecepatan studi mahasiswa menggunakan metode ID3.

Jurnal Buana Informatika, *2*(2), 102–108.

Hafidh, F., Kurniawan, M. Y., & Yazidah Anwar, R. I. (2021). Identifikasi ketunaan anak berkebutuhan khusus dengan algoritma iterative dichotomiser 3 (ID3). *Jurnal Buana Informatika*, *12*(2), 78–87.

Hikmatulloh, H., Rahmawati, A., Wintana, D., & Ambarsari, D. A. (2019). Penerapan algoritma iterative dichotomiser three (id3) dalam mendiagnosa kesehatan kehamilan. *Klik - Kumpulan Jurnal Ilmu Komputer*, *6*(2), 116.

Made, N., Juli, A., Gede, D., Divayana, H., & Indrawan, G. (2020). Analisis sentimen dokumen twitter mengenai dampak virus corona menggunakan metode naive bayes classifier. 22–29.

Makmun, A., & Radisu, I. M. (2021). Karakteristik pada Obesitas Berdasarkan Rentan Umur di Kelurahan Nganganaumala Kota Bau-Bau. *Indonesian Journal of Health*, *1*(2), 85–90.

Motif, B., Using, D., Extraction, T., Networks, A. N., Amanullah, R. F., Pujianto, A., Pratama, B. T., Informatika, M. T., Ring, J., & Utara, R. (n.d.). Deteksi motif batik menggunakan ekstraksi tekstur dan jaringan syaraf tiruan. 69–79.

Nugroho, F., Kristanto, H., & Oslan, Y. (2007). Validitas suatu alamat menggunakan decision tree dengan algoritma Id3. *Jurnal Informatika*, *3*(2), 27–23.

Putra, J. W. G. (2020). Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4 (17 Agustus 2020). 4, 45–46.

Retnoningsih, E., & Pramudita, R. (2020). Mengenal machine learning dengan

teknik supervised dan unsupervised learning menggunakan python. *Bina Insani Ict Journal*, *7*(2), 156.

Rosandy, T. (2016). Perbandingan metode naive bayes classifier dengan metode decision tree untuk menganalisa kelancaran pembiayaan. *Jurnal TIM Darmajaya*, *02*(01), 52–62.

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

Widiyoga, C. R., Saichudin, & Andiana, O. (2020). Hubungan Tingkat Pengetahuan tentang Penyakit Diabetes Melitus pada Penderita terhadap Pengaturan Pola Makan dan Physical Activity. *Sport Science Health*, *2*(2), 152–161.

Widodo, Y. B., Anggraeini, S. A., & Sutabri, T. (2021). Perancangan sistem pakar diagnosis penyakit diabetes berbasis web menggunakan algoritma naive bayes. *Jurnal Teknologi Informatika Dan Komputer*, *7*(1), 112–123.

Wijaya, H. D., & Dwiasnati, S. (2020). Implementasi data mining dengan algoritma naïve bayes pada penjualan obat. *Jurnal Informatika*, *7*(1), 1–7.

Wirawan, I. N. T., & Eksistyanto, I. (2015). Penerapan naive bayes pada intrusion detection system dengan diskritisasi variabel. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *13*(2), 182.

LAMPIRAN

Lampiran 1.

Tabel untuk prediksi klasifikasi penyakit diabetes

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	Y	pre
1	2	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1	0
2	2	1	1	1	0	1	1	0	1	1	0	1	0	1	1	1	1	1
3	3	1	0	1	1	1	1	0	1	1	1	0	0	0	1	0	1	1
4	2	1	1	1	0	0	1	1	0	1	0	1	0	1	0	0	1	1
5	3	1	1	1	0	1	1	0	1	0	0	0	1	1	0	0	1	1
6	2	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1	1
7	2	1	1	1	1	1	1	0	1	0	0	1	1	1	0	1	1	1
8	2	1	1	1	0	1	0	1	1	1	0	0	0	0	0	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
156	3	2	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1	1

lampiran pohon keputusan Algoritma *Iterative Dichotomiser Three (ID3)*