Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)
Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1

Penerapan Data Mining Untuk Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5

Zudyanti Dwi Rahma Sari[1], Jasmir Jasmir[2], Yulia Arvita[3,*]

[1] Fakultas Ilmu Komputer, Teknik Informatika, Universitas Dinamika Bangsa, Jambi, Indonesia
Email: [1]zudyanti08@gmail.com, [2]ijay_jasmir@yahoo.com, [3]yulia_arvita@yahoo.co.id
Email Penulis Korespondensi: zudyanti08@email.com

Artikel Info :
Artikel History :
Submitted : 18-02-2024
Accepted : 12-03-2024
Published : 30-04-2024

Abstrak- Kesehatan merupakan peranan terpenting dalam kehidupan. Salah satu penyakit yang dapat menyebabkan komplikasi dan kematian adalah *diabetes*. *Diabetes* merupakan penyakit yang disebabkan oleh pankreas yang tidak memproduksi insulin yang cukup untuk tubuh sehingga kadar gula dalam darah melebihi normal. *Diabetes* merupakan penyakit keturunan, penyakit ini dapat diturunkan kepada anaknya dari orang tua yang mengidap penyakit *diabetes*, sangat disayangkan jika usia yang masih muda sudah mengidap penyakit *diabetes*. Pemeriksaan dalam dunia medis dilakukan dengan cara pendiagnosaan penyakit berdasarkan gejala-gejala yang dirasakan oleh penderita yang dapat menghasilkan rekam medis gejala penyakit. Untuk meminimalisir angka kematian dari penyakit *Diabetes* ini, para pakar kesehatan harus melakukan pendiagnosaan penyakit dengan sedini mungkin. Salah satu metode yang dapat digunakan dalam permasalahan ini adalah data mining dengan teknik klasifikasi dengan menggunakan algoritma C4.5. Penelitian ini bertujuan untuk membantu para medis untuk mengklasifikasi para pasien yang memiliki gejala-gejala penyakit *diabetes*. Algoritma C4.5 merupakan metode yang digunakan untuk klasifikasi yang menghasilkan model berupa pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Algoritma C4.5 merupakan metode yang dapat digunakan untuk memprediksi dan mengetahui nilai akurasi pada pasien dengan gejala-gejala yang diderita pasien apakah pasien tersebut mengidap penyakit *Diabetes* atau tidak. Berdasarkan hasil pengujian dengan metode *cross validation* pada *tools RapidMiner* menggunakan 2 *options* yaitu *5-Fold Cross Validation* yang menghasilkan akurasi 95,88% dan *10-Fold Cross Validation* menghasilkan akurasi 95,90%. Yang mana pengujian dengan 10 *Fold Cross Validation* menghasilkan akurasi yang lebih baik dibandingkan menggunakan 5 *Fold Cross Validation*. Dengan jumlah kelas data yang sama yaitu pada *class positive* sebanyak 224 data dan pada *class negative* sebanyak 140 data.

Kata Kunci :
Diabetes, Data Mining, Prediksi, Algoritma C4.5, RapidMiner

Abstract- Diabetes is a disease caused by the pancreas not producing enough insulin for the body so that blood sugar levels exceed normal. Diabetes is a hereditary disease, this disease can be passed on to their children from parents who have diabetes, it is unfortunate if a young age already has diabetes. Examination in the medical world is carried out by diagnosing the disease based on the symptoms felt by the sufferer which can produce a medical record of the symptoms of the disease. To minimize the death rate from diabetes, health experts must diagnose the disease as early as possible. One method that can be used in this problem is data mining with classification techniques using the C4.5 algorithm. This study aims to help doctors to classify patients who have symptoms of diabetes. Algorithm C4.5 is a method used for classification which produces a model in the form of a decision tree. Decision trees are a well-known classification and prediction method. Algorithm C4.5 is a method that can be used to predict and determine the accuracy value in patients with the symptoms that the patient suffers whether the patient has diabetes or not. Based on the test results with the cross validation method on RapidMiner tools using 2 options, namely 5-Fold Cross Validation which produces 95.88% accuracy and 10-Fold Cross Validation produces 95.90% accuracy. Where testing with 10 Fold Cross Validation produces better accuracy than using 5 Fold Cross Validation. With the same number of data classes, there are 224 data in the positive class and 140 data in the negative class.

Keywords :
Diabetes, Data Mining, Prediction, C4.5 Algorithm, RapidMiner

1. PENDAHULUAN

Kesehatan merupakan peranan yang sangat penting untuk mempertahankan kehidupan manusia, dengan kesehatan yang baik manusia dapat melakukan kegiatan produktif dalam sosialisasi atau ekonomi untuk mencapai tujuan hidup mereka. Salah satu penyakit yang dapat menyebabkan komplikasi dan kematian adalah *diabetes*. *Diabetes* bukan hanya penyebab utama kematian dini di dunia, penyakit ini dapat menyebabkan kebutaan, gagal ginjal, dan bahkan penyakit jantung [1].

Diabetes adalah kondisi ketika kandungan gula atau kadar glukosa dalam darah melebihi normal yang disebabkan oleh pankreas tidak memproduksi cukup insulin untuk tubuh atau tubuh tidak menggunakan insulin

Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1

secara efektif. Insulin adalah hormon yang mengontrol gula darah, keseimbangan gula darah yang berasal dari makanan yang dimakan orang tersebar ke sel darah dalam tubuh yang dapat menghasilkan energi [2].

Diabetes disebabkan oleh kerusakan kelenjar pankreas sebagai penghasil hormon insulin dan ketidakmampuan tubuh dalam memanfaatkan insulin atau kekurangan hormon insulin sehingga kadar gula dalam darah tidak dapat terkontrol [3],[4].Gejala diabetes yang perlu kita waspadai , yaitu polydipsia (kondisi dimana merasa haus yang berlebihan), polyuria (kondisi dimana seseorang sering buang air kecil yang terjadi terutama pada malam hari), polyphagia (kondisi dimana seseorang merasa cepat lapar dan banyak makan)[5]

Di bidang medis terdapat banyak catatan pasien salah satunya adalah data penyakit diabetes. Namun, jumlah data tidak dapat digunakan dengan baik tanpa informasi dan kesimpulan dari data. Seperti sulit memprediksi apakah pasien mengidap diabetes atau tidak. Sehingga diperlukan proses ekstraksi untuk mencari informasi pada data yang sebelumnya tidak diketahui yang dikenal dengan istilah *data mining*.

Data mining adalah proses menganalisis data dari berbagai sudut dan kumpulan data yang sebelumnya tidak digunakan untuk mendapatkan pengetahuan baru dengan menemukan pola tersembunyi dalam data kemudian mengubahnya menjadi informasi yang berguna [6]. Dalam mendiagnosis diperlukan suatu metode untuk memprediksi *diabetes* secara lebih akurat dan efektif. Oleh karena itu peneliti menggunakan algoritma C4.5 untuk memprediksikan penyakit *diabetes* sehingga memberikan hasil yang akurat dari proses evaluasi.

Berdasarkan uraian dan permasalahan diatas, maka penulis melakukan penelitian yang berjudul **"Penerapan Data Mining Untuk Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5 ".** Adapun tujuan penelitian ini dilakukan oleh penulis, yaitu : Menerapkan teknik *Data Mining* dengan algoritma C4.5 dalam memprediksi penyakit *Diabetes* dan mengevaluasi hasil perhitungan algoritma C4.5 pada penderita penyakit *Diabetes* dan untuk menambah wawasan serta pengetahuan peneliti melakukan tinjauan penelitian sejenis. Berikut hasil ringkasan dari jurnal terdahulu: N. Maulidah dkk. (2021) [7], melakukan penelitian dengan judul "Prediksi Penyakit *Diabetes Mellitus* Menggunakan *Metode Support Vector Machine* dan *Naive Bayes*", Metode Support Vector Machine memiliki nilai akurasi yang jauh lebih tinggi dibandingkan dengan menggunakan metode Naive Bayes. Nilai akurasi untuk model Metode Support Vector Machine adalah 78,04% dan nilai akurasi untuk metode Naive Bayes 76,98%. Berdasarkan nilai ini, perbedaan akurasinya adalah 1,06%. F. Handayanna. Dkk. (2017) [8], melakukan penelitian dengan judul "Prediksi Penyakit *Diabetes* Menggunakan *Naive Bayes* Dengan Optimasi Parameter Menggunakan *Algoritma Genetika*", Hasil yang didapat adalah pengujian dengan menggunakan Naive Bayes didapatkan nilai accuracy adalah 72.00% sedangkan pengujian dengan menggunakan Naive Bayes berbasis Algoritma Genetika didapatkan nilai accuracy 74.74% dan Sehingga dapat disimpulkan bahwa penerapan model Naive Bayes dengan Algoritma Genetika untuk seleksi fitur dan optimalisasi parameter terbukti dapat meningkatkan akurasi dalam prediksi penyakit diabetes type 3.

Metode yang digunakan pada penelitian ini adalah metode C4.5 di mana metode ini menghasilkan pohon keputusan, yaitu algoritma klasifikasi data mining yang digunakan untuk mengubah data menjadi pohon keputusan dan aturan-aturan keputusan sehingga mendapatkan jawaban dari masalah yang dimasukan [9],[10].

Algoritma C4.5 metode yang digunakan untuk klasifikasi yang menghasilkan model berupa pohon keputusan. Algoritma ini memiliki input yaitu training samples dan samples. Dengan mengeksplorasi data yaitu menemukan hubungan variabel input atau attribute kriteria dengan variabel target atau decision attribute [11],[12].

Tahap untuk membuat pohon keputusan dengan algoritma C4.5 secara garis besar yaitu [13] :

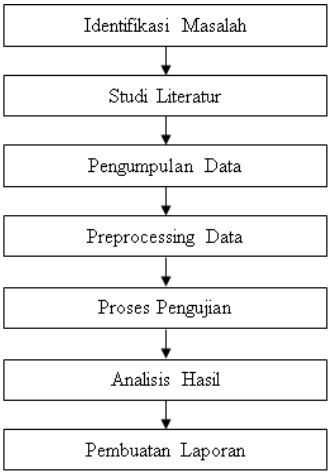
1. Tentukan nilai atribut
2. Ulangi langkah sebelumnya hingga semua kumpulan data telah dipisah
3. Proses pohon keputusan berakhir ketika:
   a. Semua perhtungan di node N diberi kelas yang sama
   b. Tidak ada atribut lain dalam kumpulan data
   c. Tidak ada catatan kosong
4. Langkah selanjutnya adalah membuat pohon yang sudah dihasilkan. Aturan diturunkan dari pohon keputusan dengan menelusuri dari daun ke daun.

Tools yang di gunakan dalam penelitian ini adalah RapidMiner, yang merupakan perangkat lunak yang berfungsi untuk alat pembelajaran dalam ilmu data mining yang mana digunakan untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi yang menggunakan berbagai macam teknik deskriptif dan prediksi sehingga dapat membuat keputusan yang paling baik [14].

2. METODOLOGI PENELITIAN

Kerangka kerja ini bertujuan untuk mempermudah memahami apa yang akan terjadi dengan penelitian dan untuk memastikan bahwa penelitian diselesaikan tepat waktu dan penelitian diselesaikan sesuai rencana. Kerangka penelitian yang digunakan adalah sebagai berikut:

Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)
Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1



Gambar 1 Kerangka Penelitian

Berdasarkan kerangka kerja penelitian yang digambarkan di atas, maka dapat di uraikan pembahasan masing-masing tahapan dalam penelitian sebagai berikut:

1. Identifikasi Masalah
   Pada tahapan ini, masalah yang diidentifikasi dalam penelitian adalah memprediksi penyakit diabetes dan mencari metode yang sesuai dengan memprediksi performa yang baik dengan menggunakan algoritma C4.5.
2. Studi Literatur
   Pada tahapan ini dilakukan pencarian landasan-landasan teori yang diperoleh dari berbagai buku, jurnal dan internet yang berhubungan dengan penelitian untuk melengkapi konsep dan teori, sehingga memiliki landasan dan keilmuan yang baik.
3. Pengumpulan Data
   Dalam penelitian ini penulis menggunakan salah satu dari dataset prediksi penyakit diabetes yang diperoleh dari website https://www.kaggle.com/ data terkait dengan penyakit diabetes dengan nama dataset Early stage diabetes risk prediction.
4. Preprocessing Data
   Tahapan preprocessing akan dilakukan tahapan sebagi berikut:
   a. *Representasi Data*
      Data penderita *diabetes mellitus* diambil dari Kaggle dengan jumlah 520 data dan 17 atribut. Variabel target *class* memiliki 2 kelas yaitu *Positive* dan *Negative*.
   b. Cleaning Data
      Tahapan ini melakukan pembersihan atau pemisahan data dari noise data dan data yang tidak konsisten. Kemudian hasil dari cleaning data akan disusun dalam bentuk csv atau arff.
   c. Transformasi Data
      Transformasi data merupakan metode yang digunakan dalam cleaning data pada tahap ini data yang telah dipisah dan dipilih lalu diubah atau diinisialkan data dari data berbentuk angka ke dalam bentuk nominal.
   d. Selection Data
      Pemilihan data baru sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam data mining dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.
   e. Data Split
      Tahapan ini dataset akan dibagi menjadi dua yaitu data training dan data testing. Data training digunakan sebagai pembentuk model dan data testing sebagai pemvalidasi model dengan perbandingan yaitu 70:30 atau 70% data training dan 30% data testing.

5. Proses Pengujian
   Pada tahap ini data akan dinilai keakuratan data yang sudah didapatkan dengan perhitungan algoritma C4.5.
   a. Klasifikasi Algoritma C4.5
      Pada tahap ini penulis melakukan proses pengklasifikasian data dengan metode Algoritma C4.5 menggunakan tool Rapid Miner untuk memprediksi penyakit diabetes terhadap data yang terpilih.
6. Interpretation / Evaluation

Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)
Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1

Prediksi atau pola informasi yang dihasilkan dari proses data mining, ditampilkan dalam bentuk analisis Algoritma C4.5 agar mudah dimengerti bagi pembaca.

7. Pembuatan Laporan
   Setelah semua tahapan penelitian dilakukan maka akan dibuat laporan sebagai dokumentasi penelitian agar dapat dimanfaatkan pada waktu yang akan datang, baik oleh peneliti sendiri, atau pun peneliti lainnya nanti.

3. HASIL DAN PEMBAHASAN

3.1 *Representasi Data*

Berdasarkan hasil pemilihan data yang sudah dilakukan, data yang akan diteliti diambil dari *dataset* (*Kaggle*) peneliti memperoleh data penderita penyakit diabetes dengan nama dataset *Early stage diabetes risk prediction* yang berjumlah 520 data dengan 17 atribut yang terdiri dari *Age*, *Gender*, *Polyuria*, *Polydipsia*, *Sudden weight lost*, *Weakness*, *Polyphagia*, *Genital thrush*, *Visual blurring*, *Itching*, *Irritability*, *Delayed healing*, *Partial paresis*, *Muscle stiffness*, *Alopecia*, *Obesity*, *Class*.

Tabel 1 Data Diabetes

No	Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Weakness	Polyphagia	Genital Thrush	Visual Blurring	Itching	Irritability	Delayed Healing	Partial Paresis	Muscle Stiffness	Alopecia	Obesity	Class
1	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
2	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
3	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
4	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
5	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
6	55	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Positive
7	57	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No	Positive
8	66	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No	No	Positive
9	67	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Positive
10	70	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	Positive
11	44	Male	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	No	Yes	Yes	No	Positive
12	38	Male	Yes	Yes	No	No	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Positive
13	35	Male	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	Positive
14	61	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Positive
15	60	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	No	No	Positive
16	58	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No	Positive
17	54	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	Yes	No	No	Positive
18	67	Male	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Positive
19	66	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Positive
20	43	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	No	No	No	No	Positive

3.2 Perhitungan Manual Algoritma C4.5

1. Entropy
   Digunakan untuk mengukur ketidakpastian dari sekumpulan atribut dari suatu set data. Ukuran *entropy* dianggap sebagai ukuran ketidakpastian dimana semakin tinggi *entropy* suatu atribut maka semakin tinggi ketidakpastian. Berikut persamaan untuk menghitung nilai entropy :
   $Entropy\ (S) = \sum_{i=1}^{n} -pi \log 2\ (pi)$ **(1)**
   Keterangan :
   S = Himpunan kasus
   n = Jumlah partisi S
   Pi = Proposisi Si terhadap S
2. Gain
   Gain merupakan informasi yang didapatkan dari perubahan entropy pada suatu kumpulan data, dengan cara melakukan partisipasi terhadap suatu set data. untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut – atribut yang ada. Untuk menghitung gain digunakan persamaan berikut :
   $Gain\ (S,A) = entropy\ (S) - \sum_{i=1}^{n} - \frac{|Si|}{S} * Entropy(Si)$ **(2)**
   Keterangan :
   S = Himpunan kasus
   A = Fitur
   n = Jumlah partisi atribut A
   |Si| = Proporsi Si terhadap S
   |S| = Jumlah kasus dalam S

Perhitungan dilakukan sesuai dengan data pada tabel 1 dimana setiap *record* dihitung untuk menentukan *entropy* dan menentukan *gain* dari setiap *record*.
Perhitungan manual menggunakan teorema persamaan 1 dan persamaan 2:

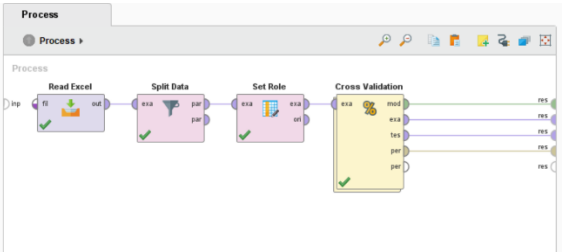
1. Menghitung *entropy* dan *gain* untuk penentuan atribut akar

Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)
Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1

a. Menghitung *entropy* total
Entropy total = ((-224/364) * log2 (224/364) + (-140/364) * log2 (140/364)) = 0,9612
b. Menghitung *entropy* dan *gain age*
Entropy age <=48 = ((-112/200) * log2 (112/200) + (-88/200) * log2 (88/200)) = 0,9896
Entropy age >=49 = ((-112/164) * log2 (112/164) + (-52/164) * log2 (52/164)) = 0,9012
Gain age = 0,9612 – (200/364 * 0,9896) + (164/364 * 0,9012) = 0,011485
c. Menghitung *entropy* dan *gain gender*
Entropy gender male = ((-104/232) * log2 (104/232) + (-128/232) * log2 (128/232)) = 0,9923
Entropy gender female = ((-120/132) * log2 (120/132) + (-12/132) * log2 (12/132)) = 0,4395
Gain gender = 0,9612 – (232/364 * 0,9923) + (132/364 * 0,4395) = 0,1694249
d. Menghitung *entropy* dan *gain polyuria*
Entropy polyuria yes = ((-171/183) * log2 (171/183) + (-12/183) * log2 (12/183)) = 0,3492
Entropy polyuria no = ((-53/181) * log2 (53/181) + (-128/181) * log2 (128/181)) = 0,8723
Gain polyuria = 0,9612 – (183/364 * 0,3492) + (181/364 * 0,8723) = 0,351915
e. Menghitung *entropy* dan *gain polydipsia*
Entropy polydipsia yes = ((-160/165) * log2 (160/165) + (-5/165) * log2 (5/165)) = 0,1959
Entropy polydipsia no = ((-64/199) * log2 (64/199) + (-135/199) * log2 (135/199)) = 0,9061
Gain polydipsia = 0,9612 – (165/364 * 0,1959) + (199/364 * 0,9061) = 0,3770519
f. Menghitung *entropy* dan *gain sudden weight loss*
Entropy sudden weight loss yes = ((-134/157) * log2 (134/157) + (-23/157) * log2 (23/157)) = 0,601
Entropy sudden weight loss no = ((-90/207) * log2 (90/207) + (-117/207) * log2 (117/207)) = 0,9877
Gain sudden weight loss = 0,9612 – (157/364 * 0,601) + (207/364 * 0,9877) = 0,1403302
g. Menghitung *entropy* dan *gain weekness*
Entropy weekness yes = ((-150/210) * log2 (150/210) + (-60/210) * log2 (60/210)) = 0,8631
Entropy weekness no = ((-74/154) * log2 (74/154) + (-80/154) * log2 (80/154)) = 0,9989
Gain weekness = 0,9612 – (210/364 * 0,8631) + (154/364 * 0,9989) = 0,0406689
h. Menghitung *entropy* dan *gain polyphagia*
Entropy polyphagia yes = ((-136/165) * log2 (136/165) + (-29/165) * log2 (29/165)) = 0,6707
Entropy polyphagia no = ((-88/199) * log2 (88/199) + (-111/199) * log2 (111/199)) = 0,9903
Gain polyphagia = 0,9612 – (165/364 * 0,6707) + (199/364 * 0,9903) = 0,1157835
i. Menghitung *entropy* dan *gain alopecia*
Entropy alopecia yes = ((-57/123) * log2 (57/123) + (-66/123) * log2 (66/123)) = 0,9961
Entropy alopecia no = ((-167/241) * log2 (167/241) + (-74/241) * log2 (74/241)) = 0,8897
Gain alopecia = 0,9612 – (123/364 * 0,9961) + (241/364 * 0,8897) = 0,0355424
j. Menghitung *entropy* dan *gain obesity*
Entropy obesity yes = ((-44/66) * log2 (44/66) + (-22/66) * log2 (22/66)) = 0,9183
Entropy obesity no = ((-180/298) * log2 (180/298) + (-118/298) * log2 (118/298)) = 0,9685
Gain obesity = 0,9612 – (66/364 * 0,9183) + (298/364 * 0,9685) = 0,0018017

3. **Pengujian Algoritma C4.5**

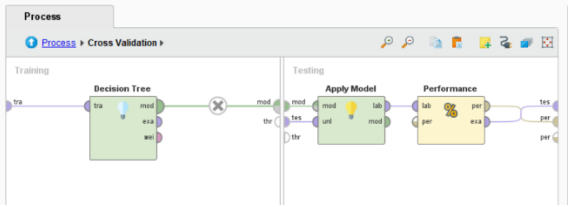
Berikut ini pengujian menggunakan *tools RapidMiner*



Gambar 1. Pengujian Algoritma C4.5

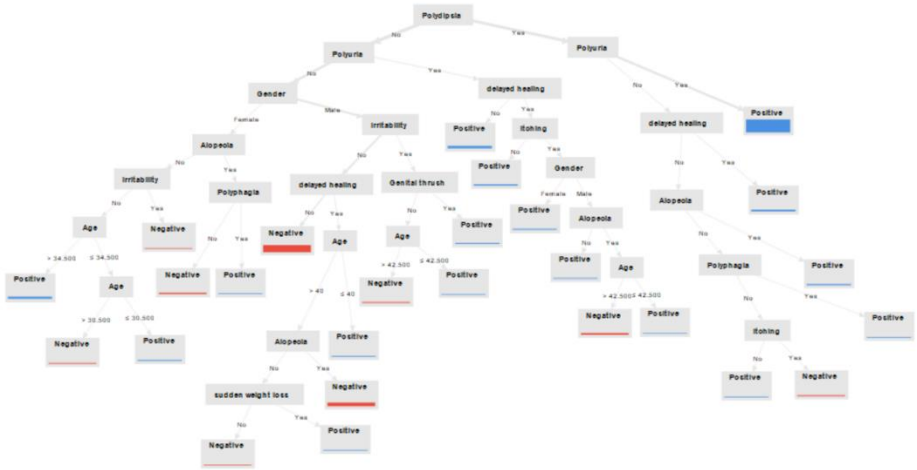
Pada menu operator cari “Read Excel” atau sesuaikan dengan jenis data yang akan di pakai. Pada bagian parameter *import* dengan mencari data yang akan di pakai. Selanjutnya pada menu operator cari “Split Data”. Pada bagian parameter *partition* kemudian masukkan jumlah rasio yang kita pakai yaitu 70% dari dataset sebagai data latih dan 30% sebagai data uji. Selanjutnya pada menu operator cari “ Set Roll”. Kemudian pada menu operator cari “ *Cross Validation”*. Pada bagian parameter *number of folds* digunakan untuk memberikan nilai k.

Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)
Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1



Gambar 2. Validation Model Algoritma C4.5

Kemudian pada bagian *training* masukkan *Decision Tree* dan pada bagian *testing* masukkan *Apply Model* dan *Performance (Clasification).* Setelah semua fungsi dihubung, lakukan run untuk proses prediksi.
Adapun hasil yang diperoleh dari pengujian bentuk graph decision tree pada gambar 4.10 dan description rule decision tree algoritma C4.5 sebagai berikut:



Gambar 3. *Graph Decision Tree Algoritma C4.5*

Berdasarkan *graph decision tree algoritma C4.5* pada gambar 4.10, maka aturan atau *rule* yang terbentuk adalah sebagai berikut:

1. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *female* dan alopecia = *no* dan irritability = *no* maka age = >34 keputusannya positive
2. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *female* dan alopecia = *no* dan irritability = *no* dan age = <=34 maka age = > 30 keputusannya negative
3. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *female* dan alopecia = *no* dan irritability = *no* dan age = <=34 maka age = <= 30 keputusannya positive
4. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *female* dan alopecia = *no* maka irritability = *yes* keputusannya positive
5. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *female* dan alopecia = *yes* maka polyphagia = *no* keputusannya negative
6. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *female* dan alopecia = *yes* maka polyphagia = *yes* keputusannya positive
7. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *male* dan irritability = *no* maka delayed healing = *no* keputusannya negative
8. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *male* dan irritability = *no* dan delayed healing = *yes* dan age = > 40 dan alopecia = *no* maka sudden weight loss = *no* keputusannya negative
9. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *male* dan irritability = *no* dan delayed healing = *yes* dan age = > 40 dan alopecia = *no* maka sudden weight loss = *yes* keputusannya positive
10. Jika *polypdisia* = *no* dan polyuria =*no* dan gender = *male* dan irritability = *no* dan delayed healing = *yes* dan age = > 40 maka alopecia = *yes* keputusannya positive
11. Jika *polypdisia* = *no* dan polyuria = *yes* dan delayed healing = *yes* dan itching = *yes* dan gender = *male* dan alopecia = *yes* dan age = <=42 maka keputusannya positive
12. Jika *polypdisia* = *yes* dan polyuria = *no* dan delayed healing = *no* dan alopecia = *no* dan polyphagia = *no* dan itching = *no* maka keputusannya positive
13. Jika *polypdisia* = *yes* dan polyuria = *no* dan delayed healing = *no* dan alopecia = *no* dan polyphagia = *no* dan itching = *yes* maka keputusannya negative

Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)
Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1

14. Jika *polypdisia* = *yes* dan polyuria = *no* dan delayed healing = *no* dan alopecia = *no* dan polyphagia = *yes* maka keputusannya positive
15. Jika *polypdisia* = *yes* dan polyuria = *no* dan delayed healing = *yes* maka keputusannya positive
16. Jika *polypdisia* = *yes* dan polyuria = *yes* maka keputusannya positive

4. **Hasil Klasifikasi Menggunakan 5-Fold Cross Validation**

accuracy: 95.88% +/- 2.37% (micro average: 95.88%)

	true Positive	true Negative	class precision
pred. Positive	214	5	97.72%
pred. Negative	10	135	93.10%
class recall	95.54%	96.43%	

Gambar 4. Nilai Akurasi Mengguunakan 5-Fold Cross Validation

Berdasarkan gambar 4. menunjukkan hasil akurasi pada *tools RapidMiner* menggunakan *test option 5-Fold Cross Validation* menunjukkan hasil akurasi sebesar 95,88% dan pada *class positive* menghasilkan *class recall* sebesar 95,54% dan *class precision* sebesar 97,72%, sedangkan *class negative* menghasilkan *class recall* sebesar 96,43% dan *class precision* sebesar 93,10%.

5. **Hasil Klasifikasi Menggunakan 10-Fold Cross Validation**

accuracy: 95.90% +/- 3.21% (micro average: 95.88%)

	true Positive	true Negative	class precision
pred. Positive	213	4	98.16%
pred. Negative	11	136	92.52%
class recall	95.09%	97.14%	

Gambar 5. Nilai Akurasi Menggunakan 10-Fold Cross Validation

Berdasarkan gambar 5. menunjukkan hasil akurasi pada *tools RapidMiner* menggunakan *test option 10-Fold Cross Validation* menunjukkan hasil akurasi sebesar 95,90% dan pada *class positive* menghasilkan *class recall* sebesar 95,09% dan *class precision* sebesar 98,16%, sedangkan *class negative* menghasilkan *class recall* sebesar 97,14% dan *class precision* sebesar 92,52%.

6. **Hasil Perbandingan Akurasi Presentasi**

Perbandingan hasil klasifikasi algoritma C4.5 dengan 2 test yaitu menggunakan 5-Fold Cross Validation dan 10-Fold Cross Validation pada RapidMiner sebagai berikut :

Tabel 2. Perbandingan Hasil Akurasi Presentasi

Model Evaluasi	**Jumlah Kelas Data**	**Akurasi**	**Presisi**	**Recall**
5 Fold Cross Validation	**224** **140**	**95,88%**	**97,72%**	**95,54%**
10 Fold Cross Validation	**224** **140**	**95,90%**	**98,16%**	**95,09%**

Berdasarkan tabel 2 yang merupakan hasil pengujian didapatkan bahwa 10 *Fold Cross Validation* menghasilkan akurasi yang lebih baik dibandingkan 5 *Fold Cross Validation* dengan hasil akurasinya yaitu 95,90% sedangkan 5 *Fold Cross Validation* menghasilkan akurasi sebesar 95,88%. Dengan jumlah kelas data yang sama yaitu pada *class positive* sebanyak 224 data dan pada *class negative* sebanyak 140 data.

Available Online at http://ejournal.unama.ac.id/index.php/jakakom
Volume 4, Nomor 1, April 2024,
ISSN 2808-5469 (media cetak), ISSN 2808-5000 (media online)
UNAMA, DOI https://doi.org/10.33998/jakakom.v4i1

4. KESIMPULAN

Kesimpulan yang dapat diambil dari hasil penelitian ini yaitu menggunakan dataset early-stage-diabetes-risk-prediction yang merupakan data yang diambil dari Kaggle yang terdiri dari 520 data dan 17 atribut.Pada proses pengujian akurasi yang dilakukan dengan tools RapidMiner menggunakan test option 5-Fold Cross Validation dan 10-Fold Cross Validation. Didapatkan bahwa hasil 10 *Fold Cross Validation* menghasilkan akurasi yang lebih baik dibandingkan 5 *Fold Cross Validation* dengan hasil akurasinya yaitu 95,90% sedangkan 5 *Fold Cross Validation* menghasilkan akurasi sebesar 95,88%. Sehingga Hasil prediksi penyakit diabetes dengan metode C4.5 dapat membantu untuk mengambil keputusan dalam memprediksi gejala yang diderita.

REFERENCES

[1] Kementerian Kesehatan RI., "Infodatin tetap produktif, cegah, dan atasi Diabetes Melitus 2020," *Pusat Data dan Informasi Kementerian Kesehatan RI*. pp. 1–10, 2020, [Online]. Available: https://pusdatin.kemkes.go.id/resources/download/pusdatin/infodatin/Infodatin-2020-Diabetes-Melitus.pdf.

[2] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.

[3] Irwan, *Epidemiologi Penyakit Tidak Menular*. Yogyakarta: Deepublish, 2016.

[4] R. Adhi Nugroho and A. Prahutama, "Klasifikasi Pasien Diabetes Mellitus Menggunakan Metode Smooth Support Vector Machine (Ssvm)," *J. Gaussian*, vol. 6, no. 3, pp. 439–448, 2017, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian.

[5] R. A. Siallagan and Fitriyani, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5," *J. Responsif Ris. Sains dan Inform.*, vol. 3, no. 1, pp. 44–52, 2021, doi: 10.51977/jti.v3i1.407.

[6] A. Muhammad and N. Muhammad, *DATA MINING Algoritma dan Implementasi*. Yogyakarta: ANDI, 2020.

[7] N. Maulidah, R. Supriyadi, D. Y. Utami, F. N. Hasan, A. Fauzi, and A. Christian, "Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes," *Indones. J. Softw. Eng.*, vol. 7, no. 1, pp. 63–68, 2021, doi: 10.31294/ijse.v7i1.10279.

[8] F. Handayanna, Rinawati, E. Arisawati, and L. S. Dewi, "Prediksi Penyakit Diabetes Menggunakan Naive Bayes Dengan Optimasi Parameter Menggunakan Algoritma Genetika," *KNiST (Konferensi Nas. Ilmu Sos. Teknol.*, pp. 71–76, 2017.

[9] R. Anief, R. Muhammad, and R. Abdul, *PENERAPAN ALGORITMA C4.5 UNTUK PREDIKSI KEPUASAN MAHASISWA TAHUN 2020*, Ke-1. Yogyakarta: Deepublish, 2021.

[10] S. Haryati, A. Sudarsono, and E. Suryana, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.

[11] F. Elfaladonna and A. Rahmadani, "Analisa Metode Classification-Decission Tree Dan Algoritma C.45 Untuk Memprediksi Penyakit Diabetes Dengan Menggunakan Aplikasi Rapid Miner," *SINTECH (Science Inf. Technol. J.*, vol. 2, no. 1, pp. 10–17, 2019, doi: 10.31598/sintechjournal.v2i1.293.

[12] Buulolo Efori, *Data Mining Untuk Perguruan Tinggi*, Ke-1. Yogyakarta: CV BUDI UTAMA, 2020.

[13] N. Azwanti, "Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada Pt. Capella Dinamik Nusantara Cabang Muka Kuning," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 13, no. 1, p. 33, 2018, doi: 10.30872/jim.v13i1.629.

[14] D. A. Pratiwi, R. M. Awangga, and M. Y. H. Setyawan, *SELEKSI CALON KELULUSAN TEPAT WAKTU MAHASISWA TEKNIK INFORMATIKA MENGGUNAKAN METODE NAIVE BAYES*. Bandung: Kreatif Industri Nusantara, 2020.