Jurnal Elektronik Ilmu Komputer Udayana
Volume 11, No 3. February 2023

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Pengaruh Metode Reduced Error Pruning pada Algoritma C4.5 untuk Prediksi Penyakit Diabetes

Luh Putu Eka Nadya Wati[a1], Ida Bagus Made Mahendra[a2], Ngurah Agus Sanjaya ER [a3], I Gusti Ngurah Anom Cahyadi Putra[a4], Agus Muliantara [a5], Luh Arida Ayu Rahning Putri [a6]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Bali, Indonesia
[1]ekanadya96@gmail.com
[2]ibm.mahendra@unud.ac.id
[3]agus_sanjaya@unud.ac.id
[4]anom.cp@unud.ac.id
[5]muliantara@unud.ac.id
[6]rahningputri@unud.ac.id

Abstract

Degenerative disease is one of the conditions that can cause the performance of several organs in the human body to decrease and affect health conditions. The prevalence rate of diabetes is predicted to continue to increase in 2030 to reach 578 million and 700 million in 2045. In this research, a diabetes prediction system was formed using the C4.5 Algorithm with Reduced Error Pruning (REP). This research is focused on the application of the Reduced Error Pruning method on the C4.5 Algorithm and used two datasets containing several medical predictors of diabetes symptoms. Based on the research that has been done, the prediction process using the C4.5 Algorithm with Reduced Error Pruning based on the first dataset resulted in an average accuracy of 92,4% with an average accuracy before Reduced Error Pruning of 91,6%. In comparison, in the second dataset, average accuracy was obtained without Reduced Error Pruning by 81,2% and 83,4% for results with Reduced Error Pruning. Based on this percentage, the Reduced Error Pruning method does not have a big influence on the level of accuracy produced.

Keywords: *Diabetes, Data Mining, Sistem Prediksi, Algoritma C4.5, Reduced Error Pruning*

1. Pendahuluan

Penyakit degeneratif merupakan salah satu kondisi yang dapat menyebabkan kinerja dari beberapa organ dalam tubuh manusia mengalami penurunan sehingga dapat mempengaruhi kondisi kesehatan. Penyakit degeneratif dapat diderita oleh semua orang dan merupakan masalah mendesak dan kontroversial bagi beberapa negara termasuk negara Indonesia. Tanpa disadari, jutaan orang memiliki kebiasaan yang tidak sehat sehingga memicu terjadinya berbagai masalah kesehatan [1]. Diabetes merupakan salah satu dari sekian banyaknya penyakit degeneratif. Banyak masyarakat mengidap penyakit ini. Berdasarkan hasil statistik dari *International Diabetes Federation* (IDF) pada tahun 2019 menyatakan bahwa Indonesia masuk ke dalam daftar negara dengan jumlah pengidap diabetes terbanyak hingga mencapai 10,7 juta kasus dan menduduki urutan ke-7 dari 10 negara. Seiring dengan pertambahan jumlah penduduk diperkirakan tingkat prevalensi penyakit diabetes akan meningkat di kalangan masyarakat pada rentang usia 65-79 tahun sebesar 19,9% atau 111,2 juta orang. Peningkatan prevalensi penyakit diabetes diprediksi akan terus bertambah pada tahun 2030 mencapai 578 juta hingga 700 juta di tahun 2045.

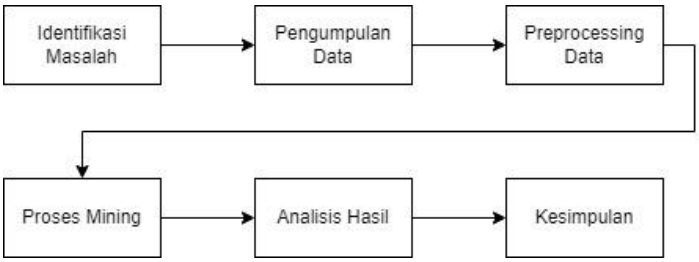
Adanya peningkatan kadar glukosa darah dalam tubuh merupakan salah satu tanda dari penyakit diabetes. Terdapat beberapa komplikasi yang bisa terjadi apabila penyakit ini tidak ditangani dengan serius. Komplikasi tersebut dapat menyebabkan adanya penurunan fungsi penglihatan yang dapat menyebabkan kebutaan, adanya penurunan fungsi ginjal, kerusakan pada beberapa saraf hingga dapat berujung pada kematian [2]. Sebanyak 25% pasien yang baru menyadari mengidap penyakit diabetes sudah mengalami komplikasi seperti neuropati sebanyak 9%, retinopati, dan nefropati sebanyak 8% berdasarkan data *United Kingdom Prospective Diabetes Study*. Pasien diabetes yang mengalami beberapa komplikasi diperkirakan telah mengidap penyakit diabetes 4-7 tahun

sebelumnya [3]. Penelitian ini akan membahas mengenai pembangunan sistem prediksi penyakit diabetes menggunakan Algoritma C4.5 dengan metode *Reduced Error Pruning*. Terdapat beberapa tahapan dalam penerapan Algoritma C4.5. Tahapan tersebut diantaranya adanya pemilihan atau pencarian atribut yang berperan sebagai *root*, pembentukan cabang untuk setiap nilai pada pohon keputusan yang dapat mendukung proses prediksi. Penerapan Algoritma C4.5 akan menghasilkan beberapa *rule* sebagai solusi permasalahan yang ingin diatasi dan mendukung proses prediksi [4]. *Reduced Error Pruning* (REP) berguna untuk meningkatkan akurasi dan menyederhanakan struktur *decision tree*. Apabila struktur *tree* yang dihasilkan sangat kompleks, hal ini dapat menyebabkan *rule* yang dihasilkan susah untuk diimplementasikan [5]. Dengan menerapkan *Reduced Error Pruning* pada Algoritma C4.5, maka akan diperoleh pengaruh REP pada tingkat akurasi sistem prediksi yang dihasilkan apabila dibandingkan dengan penerapan Algoritma C4.5 tanpa metode REP serta dapat melihat pengaruh metode REP dalam pembentukan pohon keputusan.

2. Metode Penelitian

2.1 Kerangka Penelitian

Kerangka penelitian berguna untuk mengidentifikasi tahap-tahap yang perlu dilakukan sehingga dapat memenuhi tujuan penelitian. Gambar 1 berikut ini merupakan diagram yang mencangkup beberapa tahapan kegiatan pada penelitian.



Gambar 1. Alur Penelitian

Pada kerangka penelitian diatas menjelaskan tentang tahap-tahap yang dilakukan pada penelitian ini. Identifikasi masalah merupakan tahapan awal yang dilakukan. Topik permasalahan yang diangkat pada penelitian ini yaitu perancangan sistem prediksi penyakit diabetes menggunakan Algoritma C4.5 dengan metode *Reduced Error Pruning*. Pada tahap pengumpulan data, penulis menggunakan data sekunder yang mengandung beberapa prediktor medis penyakit diabetes yang menjadi tolak ukur proses prediksi. Tahapan preprocessing data berguna untuk melihat kualitas data sebelum diproses. Tahap ini bertujuan untuk mengatasi adanya *missing value* yang terkandung pada data serta menyeleksi data-data yang ingin digunakan untuk tahap analisis sehingga dapat menghasilkan hasil yang optimal. Proses mining yang dilakukan menggunakan Algoritma C4.5 dengan *Reduced Error Pruning*. Penelitian ini membandingkan hasil proses mining Algoritma C4.5 tanpa *Reduced Error Pruning* dengan hasil proses mining Algoritma C4.5 yang ditambah dengan metode *Reduced Error Pruning*. Hasil tersebut akan dianalisis untuk mengetahui pengaruh metode *Reduced Error Pruning* terhadap Algoritma C4.5. Sehingga, pada penelitian ini dapat ditarik kesimpulan berdasarkan hasil pengujian yang dilakukan.

2.2 Penyakit Diabetes

Diabetes adalah suatu penyakit yang membuat penderitanya memiliki kadar glukosa yang sangat tinggi di dalam tubuhnya. Kurangnya produksi insulin dalam tubuh dapat menyebabkan kadar gula darah meningkat. Jika peningkatan kadar gula darah yang tinggi tidak ditangani dengan cepat maka akan berdampak pada munculnya masalah kesehatan yang lebih serius. Tingkat glukosa yang tinggi disebabkan oleh adanya kerusakan pada sel beta organ pankreas yang merupakan sumber penghasil insulin. Hal tersebut memicu ketidakefektifan hormon insulin dalam mengatur keseimbangan tingkat gula darah pada seseorang [6]. Tingkat glukosa darah yang bernilai 200 mg/dl atau lebih merupakan tanda bahwa glukosa darah tersebut tinggi. Penyakit ini dapat menyerang seluruh sistem pada tubuh seperti kulit, jantung dan hingga menimbulkan komplikasi serius. Terjadinya penyakit diabetes pada seseorang dapat dipengaruhi karena adanya resistensi insulin. Seseorang yang memiliki berat badan berlebih sangat rentan mengalami resistensi insulin. Resistensi insulin merupakan suatu keadaan yang membuat sel tubuh tidak bisa memanfaatkan hormon insulin. Insulin membantu sel tubuh menggunakan gula glukosa untuk energi. Gula darah akan menumpuk

apabila terjadi resisten terhadap insulin. Kadar gula darah akan meningkat sehingga dapat memicu hiperglikemia kronik jika produksi insulin tidak cukup kuat untuk mengkompensasi resistensi insulin. Penyebab peningkatan prevalensi diabetes adalah pertambahan populasi dan perilaku gaya hidup yang tidak sehat. Seseorang penderita penyakit ini dapat memiliki gejala antara lain adanya berkurangnya berat badan yang cukup drastis, sering merasa haus, dan memiliki rasa lapar berlebih. Selain itu, keluhan seperti badan lemah dan kurangnya energi, kesemutan di tangan atau kaki, gatal, penyembuhan luka yang lama, dan mata kabur merupakan gejala lain dari diabetes. Perubahan pola hidup menuju yang lebih sehat merupakan bentuk pencegahan dari ancaman penyakit diabetes.

2.3 Pengumpulan Dataset

Penelitian ini menggunakan dataset diabetes yang mengandung beberapa prediktor dengan dua jenis label atau kelas diabetes. Dataset yang digunakan terdiri dari dua dataset yang berbeda. Pengambilan dua dataset ini bermaksud untuk memperluas atau menambah beberapa prediktor yang menjadi tolak ukur seseorang dengan penyakit diabetes sehingga dapat meningkatkan kualitas prediksi. Dataset pertama merupakan data hasil kuesioner dari pasien rumah sakit di Sylhet, Bangladesh yang didapat dari web kaggle.com. Sedangkan untuk dataset kedua merupakan data yang didapat dari web datahub.io yang merupakan data dari Institut Nasional Diabetes dan Penyakit Pencernaan dan Ginjal. Total data prediktor medis yang digunakan terdiri dari 1.285 record data. Adapun prediktor yang akan digunakan berdasarkan dataset pertama yaitu gender, poliuria, polidipsia, penurunan berat badan secara tiba-tiba, cepat lelah, polifagia, pandangan kabur, penyembuhan tertunda, kegemukan, dan kelas. Prediktor yang digunakan pada dataset kedua yaitu kadar glukosa plasma, tekanan darah, ketebalan kulit trisep, kadar serum insulin, indeks massa tubuh, riwayat silsilah diabetes, umur serta kelas. Dataset yang digunakan termasuk ke dalam jenis data sekunder. Hal ini karena data yang digunakan diperoleh dari beberapa sumber yang tersedia di internet dan dapat diakses banyak orang. Terdapat dua kelas diabetes pada dataset yaitu positif untuk penderita penyakit diabetes dan negatif bagi yang tidak menderita penyakit diabetes.

2.4 Data Preprocessing

Data preprocessing merupakan langkah penting dalam proses data mining. Tahap ini berguna untuk meningkatkan kualitas data sebelum masuk ke tahap analisis penerapan metode [7]. Pada tahap ini dilakukan pembersihan data dengan melakukan pengisian nilai terhadap atribut yang memuat *missing value*. Data yang memiliki *missing value* akan dibersihkan dengan memberikan nilai yang frekuensinya dominan berdasarkan atribut tersebut yang nantinya akan diisi pada setiap masing-masing baris atribut yang mengandung *missing value* dengan mengacu pada atributnya. Pada dataset kedua ditemukan beberapa data *missing value* pada 5 atribut yaitu lipatan kulit trisep, insulin, glukosa plasma, tekanan darah diastolik, dan indeks massa tubuh. Pada dataset tertera bahwa terdapat nilai 0 untuk 5 atribut tersebut. Nilai 0 dianggap kurang relevan karena pada umumnya tidak mungkin seseorang memiliki indeks massa tubuh dengan nilai 0, sehingga nilai 0 pada beberapa atribut dianggap sebagai *missing value*. Data yang telah diproses kemudian akan digunakan untuk training dan testing dalam proses prediksi.

2.5 Implementasi Metode

Sistem yang telah dirancang akan diterjemahkan ke dalam bentuk kode pemrograman komputer. PHP merupakan bahasa pemrograman scripting yang digunakan pada penelitian ini untuk mengembangkan sistem prediksi dalam bentuk web. Selain itu terdapat penggunaan HTML, CSS, *framework bootstrap* untuk pembuatan tampilan website serta Javascript untuk beberapa proses validasi yang ada pada sistem. Pada penelitian ini juga memanfaatkan MySQL dalam proses pembangunan database.

2.6 Decision Tree

Decision tree adalah jenis diagram alur yang menunjukkan jalur yang jelas untuk menuju suatu keputusan dan sangat berguna untuk analisis data. Metode *decision tree* akan menghasilkan struktur pohon atau struktur hirarki keputusan yang merupakan representasi dari beberapa aturan (*rule*) sehingga mudah dipahami [8]. Klasifikasi dalam bentuk pohon keputusan dapat digunakan pada kumpulan data dengan domain medis [9]. *Decision tree* bekerja dengan mempartisi data secara rekursif berdasarkan nilai atribut input. Partisi data disebut cabang. Cabang *decision tree* akan mencakup semua parameter data yang digunakan. Selain itu, simpul daun pada *decision tree* merupakan representasi dari label suatu kelas.

2.7 Algoritma C4.5

Algoritma C4.5 digunakan dalam data mining sebagai teknik klasifikasi yang dapat digunakan untuk menghasilkan keputusan berdasarkan sampel data tertentu. Pohon keputusan yang terbentuk dapat mendukung proses pengambilan suatu keputusan. Algoritma ini memiliki beberapa kelebihan yaitu dapat menangani data numerik dan kategorik serta dapat mengolah dataset yang rumit dan besar [10]. Terdapat beberapa elemen yang harus dicari dalam pemodelan pohon keputusan menggunakan Algoritma C4.5, yaitu:

1. Entropy (S), Entropy merupakan parameter yang berguna untuk mengukur tingkat keragaman nilai atribut masing-masing terhadap suatu atribut keputusan.

   $$Entropy\ (S) = \Sigma_{i=n}^{n} - pi * log_2\ (pi) \quad (1)$$

   Keterangan:
   S = Jumlah sampel kasus (sampling)
   n = Banyaknya partisi untuk S
   pi = Perbandingan Si ke S
2. Gain (S, A) adalah nilai yang berguna sebagai dasar dalam pembuatan simpul atau akar dan cabang dari suatu pohon keputusan.

   $$Gain\ (S, A) = E(S) - \Sigma_{i=1}^{n} \frac{|Si|}{|S|} * E(Si) \quad (2)$$

   Keterangan:
   E = Entropy
   S = Jumlah sampel kasus (sampling)
   A = Atribut
   n = Banyaknya partisi untuk S
   | Si | = Banyaknya kasus pada i-partisi
   | S | = Banyaknya kasus di S

2.8 Reduced Error Pruning

Reduced Error Pruning (REP) merupakan salah satu bentuk pemangkasan dengan teknik postpruning. Postpruning yaitu proses untuk meminimalkan pohon keputusan yang terbentuk dengan cara memangkas beberapa cabang pada pohon keputusan yang sebelumnya sudah selesai dibangun. Metode REP ini merupakan metode yang bisa digunakan pada Algoritma C4.5 [5]. Cara kerja REP adalah mempertimbangkan setiap simpul atau cabang pohon untuk pemangkasan. Pemangkasan merupakan tahap menghapus subpohon dan menetapkan kelas yang paling dominan muncul di simpul itu. Sebuah simpul akan dihapus jika akurasi pohon yang dihasilkan tidak lebih buruk daripada hasil saat training serta dapat meminimalkan nilai estimasi error. Node atau cabang akan dihapus secara iteratif jika dapat meningkatkan akurasi pohon keputusan. Selain itu, proses pemangkasan dilakukan dari bagian bawah menuju bagian atas pohon. Nilai estimasi error merupakan penentu proses *pruning* perlu dilakukan atau tidak. Selama nilai estimasi error pada proses *pruning* lebih kecil atau sama dengan nilai estimasi error sebelumnya maka proses *pruning* dilakukan [11]. Proses tersebut terus dilakukan selama ditemukan nilai estimasi error yang lebih baru dan tidak mengurangi nilai akurasi yang dihasilkan dibandingkan nilai error lama. Berkaitan dengan hal ini maka terdapat proses update nilai estimasi error yang terbaru [12]. Perhitungan untuk mencari nilai estimasi error dapat dilihat sebagai berikut.

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} + \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \quad (3)$$

Keterangan:

e = Estimasi Error
f = Jumlah data yang salah terklasifikasi yang dibagi dengan jumlah data sampel.
N = Total jumlah data sampel yang digunakan.
z = Nilai konstanta yang ditetapkan Algoritma C4.5 dengan nilai 0,69.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

2.9 Desain Evaluasi Metode

Pengujian pada penelitian ini dilakukan untuk mengetahui tingkat akurasi serta nilai estimasi error berdasarkan proses prediksi. Pada umumnya untuk mengevaluasi kinerja algoritma klasifikasi menggunakan skenario pengujian *K-Fold Cross Validation*. Pengujian menggunakan *K-Fold Cross Validation* berguna untuk mengetahui hasil rata-rata akurasi model sebanyak k pengujian dengan memprediksi beberapa kumpulan data yang diinput secara acak. Pada *K-Fold Cross Validation* terdapat pembagian dataset dengan rasio yang sama sebanyak nilai k [13]. Masing-masing dari *k-fold* diberikan kesempatan untuk digunakan sebagai data testing, sementara semua *fold* lainnya secara kolektif digunakan sebagai data training. Pada penelitian ini menggunakan jumlah *fold* sebanyak 5 untuk menguji validitas dalam model. Penelitian ini menggunakan perbandingan data sebesar 80% data untuk proses training dan 20% data digunakan pada proses testing. Hal ini juga dapat memastikan bahwa sistem akan menghasilkan keluaran sesuai dengan yang diinginkan berdasarkan masukan tertentu.

Setelah dilakukannya penerapan Algoritma C4.5 dengan *Reduce Error Pruning* (REP), maka selanjutnya dilakukan proses untuk menghitung nilai akurasi berdasarkan pada data yang berhasil diprediksi secara benar, yaitu dengan menjumlahkan data yang diprediksi dengan benar oleh sistem yang kemudian dibagi dengan jumlah total data yang digunakan dan dikalikan dengan 100%. Berikut ini merupakan rumus perhitungan akurasi dari sistem prediksi pada penelitian ini [14], yaitu:

$$\text{Akurasi} = \frac{\sum klasifikasi\ benar}{\sum data\ uji} \times 100\% \quad (4)$$

Keterangan :

Akurasi = hasil tingkat akurasi
klasifikasi benar = banyaknya data yang berhasil prediksi secara benar
data uji = total keseluruhan data uji yang digunakan

3. Hasil dan Pembahasan

Berdasarkan pada keseluruhan alur kerangka penelitian yang telah dilakukan. Maka untuk hasil penelitian dapat dilihat pada bagian berikut ini.

3.1 Pengujian Akurasi

Percobaan dilakukan beberapa kali dengan skenario *K-Fold Cross Validation* dengan menentukan nilai k terlebih dahulu sebanyak 5. Pada penelitian ini menggunakan dua dataset sehingga terdapat dua tabel hasil akurasi menggunakan *Reduced Error Pruning* berdasarkan dataset. Jumlah rasio terkait data yang berguna untuk proses training dan data uji pada setiap percobaan yaitu sebesar 80% untuk data training dan 20% berguna sebagai data uji. Berdasarkan persentase tersebut maka untuk dataset pertama jumlah data training yang digunakan sebanyak 416 dan 104 sebagai data uji. Sedangkan untuk dataset kedua menggunakan data sebanyak 612 untuk data training dan 153 digunakan sebagai data uji.

3.1.1 Pengujian Dataset Pertama

Hasil pengujian yang sudah dilakukan pada dataset pertama dapat diamati pada Tabel 1 dan Tabel 2 berikut ini.

Tabel 1. Pengujian Dataset 1 tanpa *Reduced Error Pruning*

Pengujian ke-	Banyaknya Prediksi Benar	Banyaknya Prediksi Salah	Akurasi	Estimasi Error
1	92	12	88 %	12 %
2	97	7	93 %	7 %
3	96	8	92%	8 %

4	95	9	91 %	9 %
5	98	6	94 %	6 %
	Rata-rata		91,6%	8,4%

Tabel 2. Pengujian Dataset 1 dengan *Reduced Error Pruning*

Pengujian ke-	Banyaknya Prediksi Benar	Banyaknya Prediksi Salah	Akurasi	Estimasi Error
1	93	11	89 %	11 %
2	98	6	94 %	6 %
3	96	8	92%	8 %
4	95	9	91 %	9 %
5	100	4	96 %	4 %
	Rata-rata		92,4%	7,6%

Tabel 1 menunjukkan hasil dari pengujian sistem prediksi menggunakan Algoritma C4.5 tanpa metode *Reduced Error Pruning* berdasarkan prediktor medis pada dataset pertama. Pada pengujian tersebut yang terlihat pada Tabel 3 menghasilkan rata-rata akurasi sebesar 91,6 %. Berdasarkan pada Tabel 2 menunjukkan hasil dari pengujian sistem prediksi pada dataset pertama menggunakan Algoritma C4.5 dengan *Reduced Error Pruning*. Hasil rata-rata akurasi secara keseluruhan pada skenario pengujian menggunakan *K-Fold Cross Validation* dengan *fold* sebanyak 5 sebesar 92,4 %. Mengacu pada hasil tersebut maka perolehan nilai akurasi dengan penerapan *Reduced Error Pruning* mengalami peningkatan yang tidak banyak yaitu sebesar 0,8%.

3.1.2 Pengujian Dataset Kedua

Pada proses implementasi Algoritma C4.5 untuk dataset kedua dilakukan pemisahan nilai ambang batas dikarenakan seluruh atribut yang terkandung pada dataset kedua bersifat numerik. Pemisahan nilai numerik dilakukan dengan mencoba setiap nilai sebagai titik ambang dan menghitung perolehan akurasi untuk setiap nilai ambang batas [15]. Hal ini berguna untuk melihat nilai ambang batas yang dapat menghasilkan model prediksi terbaik. Berikut ini Tabel 3 merupakan penentuan nilai ambang batas yang digunakan untuk penanganan data numerik berdasarkan pada dataset kedua.

Tabel 3. Penentuan Nilai Ambang Batas

No.	Nama Field	Nilai Ambang Batas
1.	Kadar Glukosa Darah	<= 140 dan > 140
		<= 126 dan > 126
		<= 99 dan > 99
		<= 70 dan > 70
2.	Tekanan Darah	<= 110 dan > 110
		<= 90 dan > 90
		<= 80 dan > 80
3.	Ketebalan Lipatan Kulit Trisep	<= 35 dan > 35
4.	Kadar Serum Insulin	<= 750 dan > 750
		<= 450 dan > 450
		<= 150 dan > 150
5.	Indeks Massa Tubuh	<= 30,1 dan > 30,1
		<= 23,4 dan > 23,4
		<= 18,5 dan > 18,5

6.	Riwayat Silsilah Diabetes	<= 0,56 dan > 0,56 <= 0,39 dan > 0,39
7.	Umur	<= 61 dan > 61 <= 45 dan > 45 <= 28 dan > 28 <= 25 dan > 25

Berikut ini pada Tabel 4 dan Tabel 5 merupakan hasil pengujian berdasarkan nilai ambang batas yang digunakan untuk menangani data numerik pada dataset kedua, seperti yang tertera pada Tabel 3.

Tabel 4. Pengujian Kedua Dataset 2 tanpa *Reduced Error Pruning*

Pengujian tahap-	Banyaknya Prediksi Benar	Banyaknya Prediksi Salah	Akurasi	Estimasi Error
1	123	30	80 %	20 %
2	129	24	84 %	16 %
3	127	26	83 %	17 %
4	120	33	78 %	22 %
5	124	29	81 %	19 %
Rata-rata			81,2%	18,8%

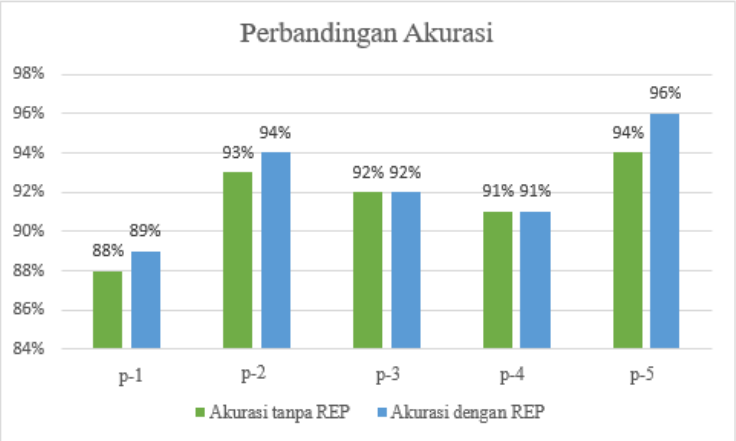
Tabel 5. Pengujian Kedua Dataset 2 dengan *Reduced Error Pruning*

Pengujian tahap-	Banyaknya Prediksi Benar	Banyaknya Prediksi Salah	Akurasi	Estimasi Error
1	125	28	82 %	18 %
2	131	22	86 %	14 %
3	130	23	85 %	15 %
4	124	29	81 %	19 %
5	127	26	83 %	17 %
Rata-rata			83,4%	16,6%

Tabel 5 berikut ini merupakan tabel hasil uji Algoritma C4.5 dengan *Reduced Error Pruning* pada dataset kedua. Dari hasil tersebut dapat dilihat bahwa persentase akurasi yang dihasilkan cukup baik jika dibandingkan dengan hasil pada Tabel 4. Berdasarkan hasil diatas maka didapatkan rata-rata akurasi dengan *Reduced Error Pruning* dari keseluruhan percobaan dengan beberapa nilai ambang batas pada dataset kedua sebesar 83,4 %. Nilai akurasi meningkat sebesar 2,2% menggunakan metode *Reduced Error Pruning*.

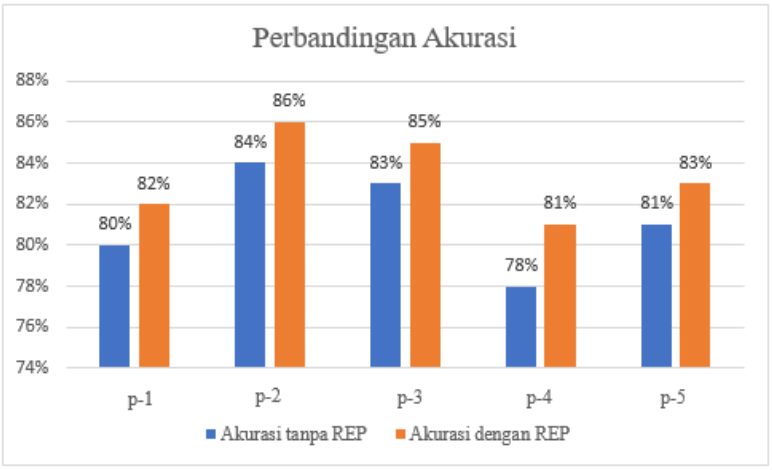
3.1 Pengaruh Metode Reduced Error Pruning

Pada penelitian ini, pengujian dilakukan dengan membandingkan hasil Algoritma C4.5 tanpa *Reduced Error Pruning* dan hasil dengan *Reduced Error Pruning* untuk mengetahui seberapa besar pengaruh penerapan *Reduced Error Pruning*. Proses eksperimen ini membandingkan akurasi serta jumlah *rule* yang dihasilkan. Pengaruh penerapan metode *Reduced Error Pruning* terhadap besar akurasi yang dihasilkan dapat dilihat pada grafik berikut yang tercantum pada Gambar 2 serta Gambar 3.



Gambar 2. Perbandingan Akurasi Dataset 1

Berdasarkan Gambar 2 tersebut menunjukkan bahwa metode *Reduced Error Pruning* tidak memberikan pengaruh terhadap peningkatan akurasi yang dihasilkan berdasarkan dataset pertama. Hal tersebut dapat dilihat dari hasil pengujian ke-3 dan ke-4 yang menghasilkan akurasi sama dengan hasil akurasi tanpa *Reduced Error Pruning*. Sementara itu, jika dilihat pada hasil pengujian ke-1, ke-2 dan ke-5 penerapan *Reduced Error Pruning* memiliki pengaruh terhadap akurasi yang dihasilkan hal tersebut dilihat dari adanya peningkatan nilai akurasi dari sebelumnya. Pada pengujian ke-1 dan ke-2 nilai akurasi naik sebesar 1% sedangkan pada pengujian ke-5 nilai akurasi naik sebesar 2% menggunakan *Reduced Error Pruning*.

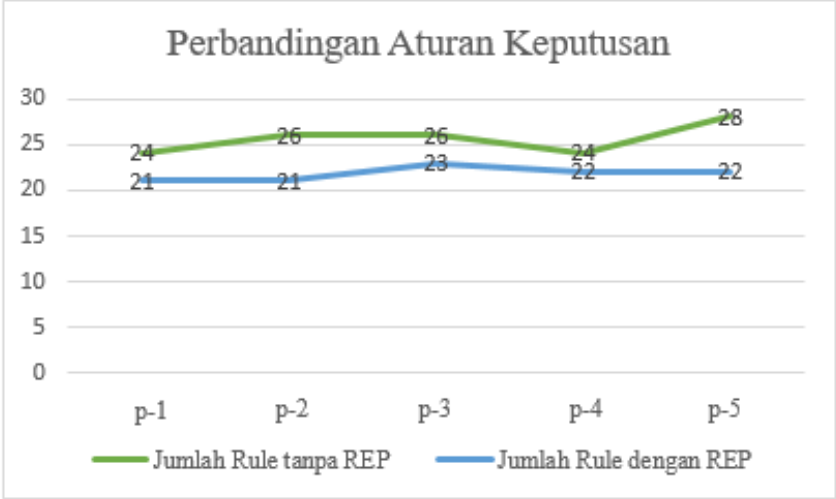


Gambar 3. Perbandingan Akurasi Dataset 2

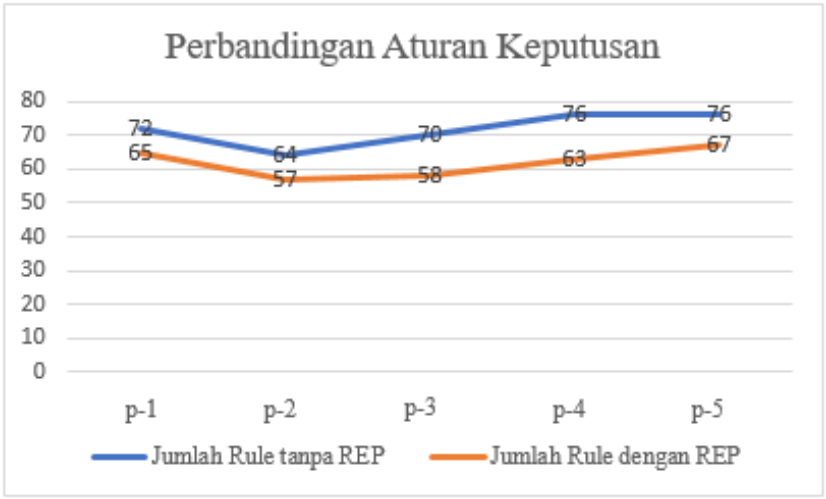
Berbeda pada hasil pengujian pada dataset pertama, perbandingan akurasi berdasarkan pada pengujian yang dilakukan pada dataset kedua menunjukkan bahwa metode *Reduced Error Pruning* dapat meningkatkan akurasi pada setiap percobaan yang dilakukan. Berdasarkan pada Gambar 3 terlihat bahwa peningkatan akurasi dengan metode *Reduced Error Pruning* pada setiap percobaan tidak menghasilkan nilai akurasi yang jauh berbeda dari sebelumnya jika dibandingkan dengan hasil tanpa *Reduced Error Pruning*. Adanya perbandingan akurasi antara percobaan pada dataset pertama dan kedua, maka dapat diketahui bahwa penerapan metode *Reduce Error Pruning* pada Algoritma C4.5 tidak selalu memberikan pengaruh terhadap nilai akurasi yang dihasilkan pada setiap percobaan.

Selain perbandingan hasil akurasi, penelitian ini juga membandingkan jumlah aturan keputusan yang berhasil dibentuk menggunakan Algoritma C4.5 tanpa *Reduced Error Pruning* maupun dengan *Reduced Error Pruning*.

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 4. Perbandingan Aturan Keputusan Dataset 1



Gambar 5. Perbandingan Aturan Keputusan Dataset 2

Berdasarkan grafik yang tertera pada Gambar 4 dan Gambar 5, terlihat bahwa baik untuk dataset pertama maupun dataset kedua yang diuji dengan menggunakan metode *Reduced Error Pruning* menghasilkan aturan keputusan yang lebih sedikit jika dibandingkan dengan aturan keputusan yang dihasilkan tanpa adanya *Reduced Error Pruning*. Jika dilihat dari hasil percobaan tersebut maka dapat disebutkan bahwa jumlah *rule* yang berhasil terbentuk menggunakan Algoritma C4.5 dapat disederhanakan melalui penerapan metode *Reduced Error Pruning*. Hal ini dikarenakan jumlah aturan keputusan yang terbentuk mengalami penurunan sebesar 14,58% untuk pengujian dataset pertama dan 13,32% untuk dataset kedua berdasarkan rata-rata persentase penurunan jumlah aturan keputusan antara penerapan metode Algoritma C4.5 tanpa adanya *Reduced Error Pruning* dibandingkan dengan adanya *Reduced Error Pruning*.

4. Kesimpulan

Berdasarkan hasil uji coba proses prediksi penyakit diabetes menggunakan Algoritma C4.5 dengan *Reduced Error Pruning* berdasarkan pada dataset pertama dihasilkan rata-rata akurasi sebesar 92,4% dengan rata-rata akurasi sebelum *Reduced Error Pruning* sebesar 91,6% sedangkan pada dataset kedua dihasilkan rata-rata akurasi tanpa *Reduced Error Pruning* sebesar 81,2% dan 83,4% untuk hasil dengan *Reduced Error Pruning*. Tingkat akurasi menggunakan metode *Reduced Error Pruning* mengalami peningkatan sebesar 0,8% untuk dataset pertama dan 2,2% untuk dataset kedua. Pada penelitian ini metode *Reduced Error Pruning* memiliki pengaruh terhadap jumlah aturan keputusan yang terbentuk. Jumlah aturan keputusan yang dihasilkan dengan *Reduced Error Pruning* lebih sedikit dengan rata-rata penurunan sebesar 14,58% untuk pengujian dataset pertama dan

13,32% untuk dataset kedua dibandingkan dengan aturan keputusan yang terbentuk melalui Algoritma C4.5 tanpa *Reduced Error Pruning*.

References

[1] I. Istianah, S. Septiani, and G. K. Dewi, "Mengidentifikasi Faktor Gizi pada Pasien Diabetes Mellitus Tipe 2 di Kota Depok Tahun 2019," *J. Kesehat. Indones. (The Indones. J. Heal.*, vol. X, no. 2, pp. 72–78, 2020.

[2] I. Irma, L. O. Alifariki, and A. Kusnan, "Uji Sensitifitas dan Spesifisitas Keluhan Penderita Diabetes Melitus Berdasarkan Keluhan dan Hasil Pemeriksaan Gula Darah Sewaktu (GDS)," *J. Kedokt. dan Kesehat.*, vol. 16, no. 1, p. 25, 2020, doi: 10.24853/jkk.16.1.25-34.

[3] "Issn: 2089-9084 ism, vol. 6 no.1, mei - agustus," vol. 6, no. 1, 2015.

[4] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti," *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017, doi: 10.15575/join.v2i1.71.

[5] R. K. Amin, Indwiarti, and Y. Sibaroni, "Implementasi Klasifikasi Decision Tree Dengan Algoritma C4 . 5 Dalam Pengambilan Keputusan Permohonan Kredit Oleh Debitur," *e-Proceeding Eng.*, vol. 2, no. 1, 2015.

[6] W. Wijanarto and R. Puspitasari, "Optimasi Algoritma Klasifikasi Biner dengan Tuning Parameter pada Penyakit Diabetes Mellitus," *Eksplora Inform.*, vol. 9, no. 1, pp. 50–59, 2019, doi: 10.30864/eksplora.v9i1.257.

[7] A. G. Lazuardy, H. S. S. Kom, and M. Eng, "Data Cleansing Pada Data Rumah Sakit," pp. 1–6, 2019.

[8] M. Yunus, H. Ramadhan, D. R. Aji, and A. Yulianto, "Penerapan Metode Data Mining C4.5 Untuk Pemilihan Penerima Kartu Indonesia Pintar (KIP)," *Paradig. - J. Komput. dan Inform.*, vol. 23, no. 2, 2021, doi: 10.31294/p.v23i2.11395.

[9] A. Kumar and B. K. Sarkar, "A hybrid predictive model integrating C4.5 and decision table classifiers for medical data sets," *J. Inf. Technol. Res.*, vol. 11, no. 2, pp. 150–167, 2018, doi: 10.4018/JITR.2018040109.

[10] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," *Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 1937–1946, 2014, doi: 10.1016/j.eswa.2013.08.089.

[11] Y. Kustiyahningsih and E. Rahmanita, "Aplikasi Sistem Pendukung Keputusan Menggunakan Algoritma C4.5. untuk Penjurusan SMA," *J. Semantec*, vol. 5, no. 2, pp. 101–108, 2016.

[12] I. Iskandar *et al.*, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4 . 5 Dengan Teknik Pruning," *J. Ilmu Komput. dan Sist. Inf.*, vol. 6, no. 1, pp. 64–68, 2018.

[13] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[14] A. Prasatya, R. R. A. Siregar, and R. Arianto, "Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes," *Petir*, vol. 13, no. 1, pp. 86–100, 2020, doi: 10.33322/petir.v13i1.925.

[15] D. Putra and I. G. A. G. A. Kadnyanana, "Implementation of Feature Selection using Information Gain Algorithm and Discretization with NSL-KDD Intrusion Detection System," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 9, no. 3, p. 359, 2021, doi: 10.24843/jlk.2021.v09.i03.p06.