PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika
Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes

Andika Prasatya[1]; Riki Ruli A Siregar[2]; Rakhmat Arianto[3]

[1, 2]Program Studi Teknik Informatika, Institut Teknologi PLN
[3]Politeknik Negeri Malang
[1]andika1531220@itpln.ac.id
[2]riki.ruli@itpln.ac.id
[3]ari87anto@gmail.com

ABSTRACT

The purpose of this study is to predict HbA1c in diabetics. The obstacles behind the prediction of HbA1c is the limitations in the laboratory to provide services for diabetics regarding HbA1 check-up. HbA1c prediction is made by a combination of K-Means and C4.5 methods. K-Means is used to classify continuous data. From the results of the K-Means classification will be used by C4.5 to create a rule (decision tree). The prediction results obtained will be carried out as a validation process to determine the level of accuracy by using K-Fold Cross-Validation. The accuracy value obtained is 72%. The resulting benefit from the prediction of HbA1c can be used as an alternative solution to overcome limitations in the laboratory in terms of HbA1c check-up servicing and the results of HbA1c prediction can also be used as a recommendation by doctor in determining the medical decision for diabetics.

Keywords: *Combinations, Predictions, HbA1c, Validation, Diabetics*

ABSTRAK

Tujuan dari penelitian ini yaitu melakukan prediksi HbA1c pada penderita diabetes. Adapun kendala yang melatarbelakangi prediksi HbA1c adalah adanya keterbatasan pada laboratorium untuk memberikan pelayanan kepada penderita diabetes dalam hal pemeriksaan HbA1c. Prediksi HbA1c dilakukan dengan kombinasi metode K-Means dengan C4.5. K-Means digunakan untuk mengelompokkan data yang bersifat kontinu. Dari hasil pengelompokan K-Means akan digunakan oleh C4.5 untuk membuat rule (pohon keputusan). Hasil prediksi yang didapatkan akan dilakukan proses validasi untuk mengetahui tingkat keakurasian dengan menggunakan K-Fold Cross Validation. Nilai akurasi yang didapatkan sebesar 72%. Manfaat yang dihasilkan dari prediksi HbA1c adalah dapat digunakan sebagai alternatif solusi untuk mengatasi keterbatasan pada laboratorium dalam hal pelayanan pemeriksaan HbA1c dan hasil prediksi HbA1c dapat juga digunakan sebagai rekomendasi oleh dokter dalam menentukan keputusan medis pada penderita diabetes.

Kata kunci: *Kombinasi, Prediksi, HbA1c, Validasi, Diabetes*

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

1. PENDAHULUAN

Hasil survei mengenai masyarakat yang menderita diabetes, didapatkan sekitar 30% penderita diabetes tidak mengetahui penyakitnya dan penderita diabetes tersebut baru mengetahuinya setelah mendapatkan hasil diagnosis keluar dari hasil tes laboratorium, bahkan sekitar 25% telah terkena komplikasi *mikrovaskular* [1]. Komplikasi *mikrovaskular* yakni komplikasi yang dapat menyebabkan kerusakan pada mata, ginjal, indra perasa, dan lain-lain yang terjadi pada pembuluh darah kecil. Diprediksi ada 439 juta orang yang pada tahun 2030 akan menderita penyakit diabetes. Sehingga penyakit diabetes merupakan persoalan kesehatan di semua negara [2].

Hemoglobin yang berhubungan dengan glukosa disebut dengan *HbA1c* (hemoglobin A1c) atau *glycated* hemoglobin. Glukosa akan saling mengikat dengan hemoglobin yang ada didalam sel darah merah yang terjadi didalam darah. HbA1c ini nanti digunakan dokter untuk memberikan keputusan yang sesuai bagi penderita diabetes. Keputusan tersebut mengandung obat yang akan diberikan, larangan yang diberikan untuk penderita, berapa kali harus menemui dokter untuk mengontrol diabetes penderita dan lain-lain. Usulan diperlukan prediksi *HbA1c* untuk mengatasi keterbatasan dalam pemberian pelayanan tes *HbA1c* pada laboratorium bagi penderita diabetes.

Adapun penelitian yang berhubungan dengan bidang medis yang dimana pada penelitian tersebut, membandingkan algoritma *Decision Tree C4.5* dengan algoritma *Naive Bayes*. Hasil akurasi yang diperoleh yaitu algoritma *Decision Tree* C4.5 sebesar 90%. Untuk algoritma *Naive Bayes* sebesar 89.58% [3]. Dan terdapat juga hasil penelitian mengenai pengelompokan data yang dimana penelitian tersebut membandingkan metode *Clustering K-Means* dengan *Agglomerative Hierarchical Clustering* (AHC). Hasil yang didapatkan dari penelitian tersebut menunjukkan bahwa metode *Clustering K-Means* lebih baik dalam mengelompokkan data dibandingkan dengan metode *Agglomerative Hierarchical Clustering* (*AHC*) [4].

Sehingga dapat disimpulkan dalam penelitian ini akan menerapkan metode klasifikasi C4.5 untuk memprediksi *HbA1C* pasien penderita diabetes. Dan untuk pengelompokan menggunakan metode *clustering K-Means*. Pengelompokan dilakukan terlebih dahulu sebelum melakukan proses metode *C4.5*.

2. METODE PENELITIAN

Tahapan-tahapan yang digunakan dalam proses data *mining* adalah CRISP-DM (Cross-Industry Standard Process Model for Data Mining). Adapun gambaran tahapan diagram CRISP-DM sebagai berikut (gambar 1):

Gambar 1. Diagram CRISP-DM [5]

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

2.1. *Business Understanding*

Menjelaskan bagaimana bisnis atau sistem yang sedang berjalan yang berkaitan dengan penelitian yang dilakukan. Adapun penelitian pada tahapan *business understanding* sebagai berikut:

a. Menentukan Tujuan Bisnis (*Determine Business Objectives*)
   Untuk mengenali proses diagnosis diabetes pada pasien dan mengetahui proses penanganan pasien penderita diabetes.
b. Menilai Situasi (*Assess the Situation*)
   Adapun diagram alur proses diagnosis diabetes pada pasien dan diagram alur perawatan pasien yang telah terkena diabetes [6].
c. Menentukan Tujuan Data Mining (*Determine the Data Mining Goals*)
   Tujuan data *mining* adalah untuk memprediksikan hasil *HbA1c*.

2.2. Data Understanding

Memahami data yang akan digunakan dalam penelitian. Adapun penelitian pada tahapan *data understanding* sebagai berikut:

a. Mengumpulkan Data Awal (*Collect the Initial Data*)
   Data yang dikumpulkan dari *dataset* yang diambil dari UCI *Machine Learning* yang di publikasi kan dari penelitian sebelumnya yang berjudul *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database* [7].
b. Mendeskripsikan Data (*Describe the Data*)
   Mendeskripsi setiap atribut–atribut yang terdapat pada *dataset* yang digunakan.
c. Mengeksplorasi Data (*Explore the Data*)
   Mendeskripsikan tipe nilai dan jenis nilai pada tiap atribut pada *dataset* yang digunakan.
d. Memverifikasi Kualitas Data (*Verify Data Quality*)
   Mendeskripsikan total nilai kosong (*missing*) pada tiap atribut pada *dataset* yang digunakan.

2.3. Data Preparation

a. Mendeskripsikan Data Set (*Data* Set *Description*)
   Deskripsi data sesuai pada 2.2. bagian b. mendeskripsikan data.
b. Memilih Data (Select Data)
   Pada penelitian lebih bertujuan pada memprediksi nilai HbA1c. Untuk hal tersebut dari atribut–atribut data pada *dataset* hanya akan di digunakan beberapa yang dibutuhkan dalam proses prediksi HbA1c. Adapun atribut–atribut data yang digunakan untuk metode *K-Means* dan *C4.5* sebagai berikut:

Tabel 1. Atribut Metode K-Means

No	Nama Atribut
1	*Time in Hospital*
2	*Numbers of Lab Procedures*
3	*Numbers of Procedures*
4	*Number of Diagnoses*

Tabel 2. Atribut Metode C4.5

No	Nama Atribut
1	*Gender*
2	*Age*
3	*Admission Type*
4	*Discharge Disposition*

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

No	Nama Atribut
5	*Admission Source*
6	*Diagnosis 1*
7	*Diagnosis 2*
8	*Diagnosis 3*
9	*A1c test result*
10	*Change of Medications*
11	*Diabetes Medications*
12	*Readmitted*

Atribut hasil proses *K-Means* akan digunakan pada proses C4.5. Pada tabel 1 dan 2 menjelaskan atribut nilai yang akan digunakan langsung berdasarkan dari *dataset* yang digunakan.

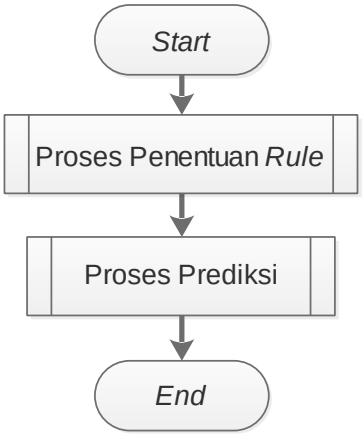
c. Membersihkan Data (*Clean Data*)

Dari hasil atribut–atribut yang telah dipilih, akan pembersihan pada data yang terdapat pada atribut ini. Dalam penelitian ini akan menghapus data dari atribut A1c *Test Result* yang bernilai *None*. Tujuan menghapus nilai adalah agar proses klasifikasi dengan metode C4.5 hasil yang dikeluarkan lebih tepat dan hasil prediksi tidak terdapat nilai *none* (tidak diketahui).

Adapun jumlah data dalam data set tersebut berjumlah 101767 data. Setelah dilakukan pembersihan data, data yang tersisa berjumlah 17019.

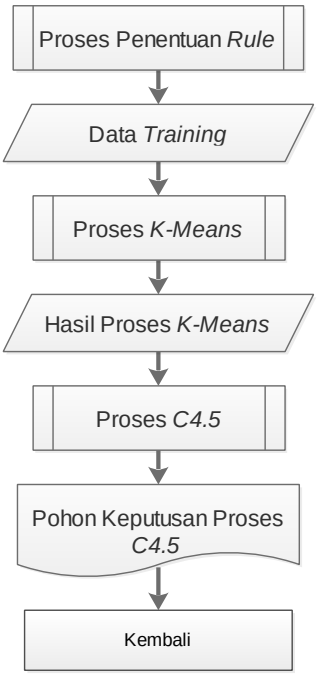
2.4. Modeling

Modeling pada tahapan ini menjelaskan bagaimana pemodelan yang akan diterapkan dalam data *mining*. Adapun gambaran langkah-langkah pemodelan data *mining* sebagai berikut:

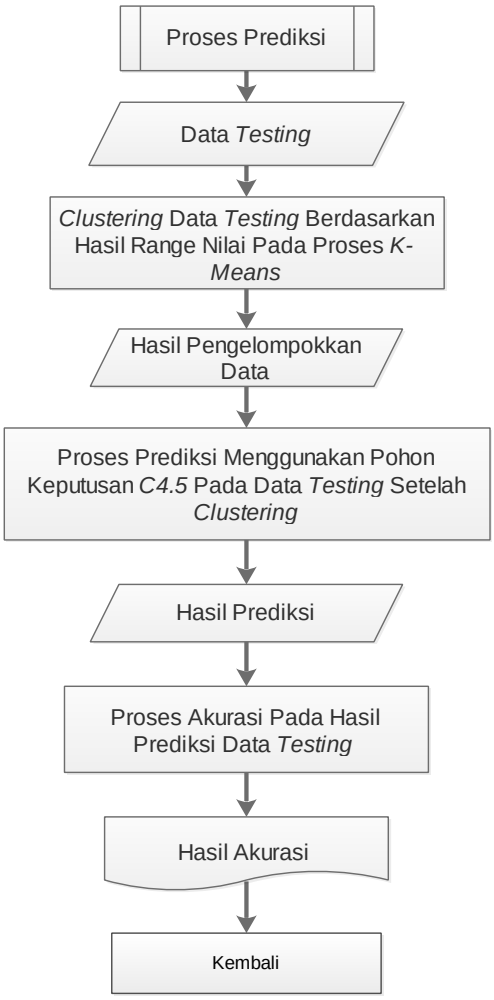


Gambar 2. Alur Pemodelan Utama

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

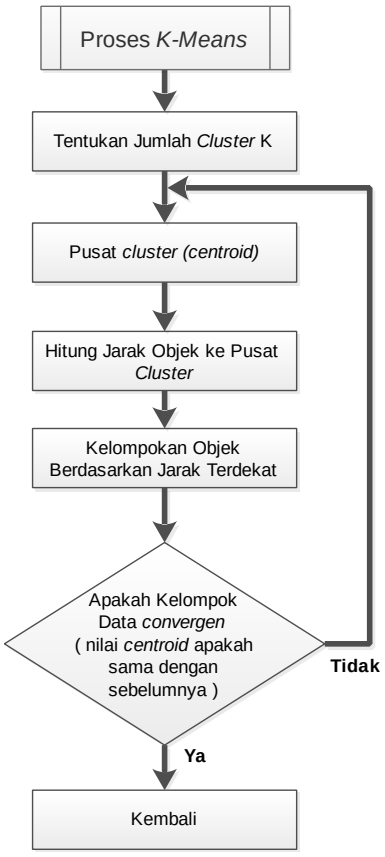


Gambar 3. Alur Pemodelan Pohon Keputusan (Rule) Prediksi HbA1c

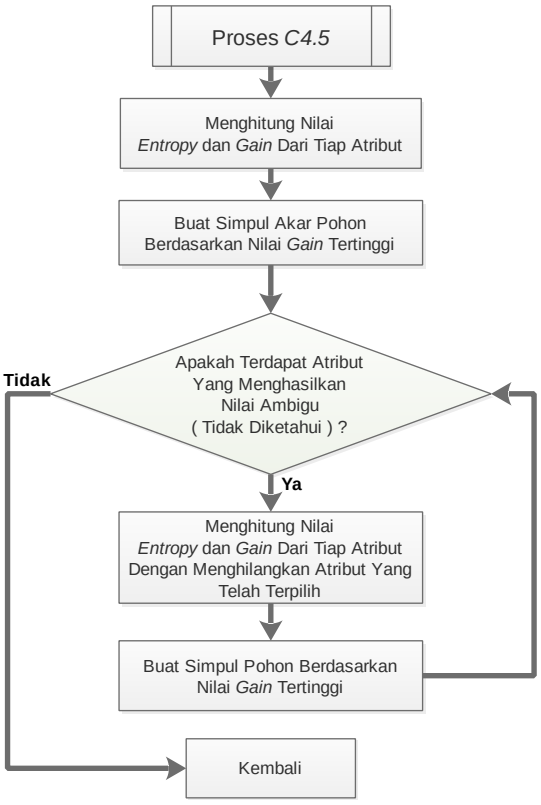


Gambar 4. Alur Pemodelan Prediksi HbA1c

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925



Gambar 5. Diagram Tahapan *K-Means* [8]



Gambar 6. Diagram Tahapan *C4.5* [9]

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

Selanjutnya adalah rumus yang akan digunakan dari tiap metode sebagai berikut :

a. Metode *K-Means*

1. Rumus Mencari Jarak Objek Centroid (*Euclidean Distance*) [8].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_{i-}y_i)^2} \qquad (1)$$

Keterangan:
d = jarak antara x dan y
x = data pusat klaster
y = data pada atribut
i = setiap data
n = jumlah data
xi = data pada pusat klaster ke i
yi = data pada setiap data ke i

2. Rumus Menentukan *Centroid* Baru [8].

$$C_{i\,=} \frac{\sum_{i=1}^{n} x_i \in s_i}{n} \qquad (2)$$

Keterangan:
Ci = centroid baru ke i
si = objek ke i
xi = nilai pada objek ke i
n = jumlah data pada tiap kelompok

b. Metode *C4.5*

1. Rumus Entropy [10]

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log 2\ pi \qquad (3)$$

Keterangan:
S = Himpunan Kasus
n = jumlah partisi S
pi = proporsi dari Si terhadap S

2. Rumus Gain [10]

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \qquad (4)$$

Keterangan:
S = Himpunan Kasus
A = Atribut
n = jumlah partisi atribut A
|Si| = jumlah kasus pada partisi ke-i
|S| = jumlah kasus dalam S

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

Adapun penerapan proses metode *K-Means* dan *C4.5* sebagai berikut:

A. Proses *K-Means*

Pada proses *K-Means* hanya akan dijelaskan sebagian. Contoh proses yang dijelaskan disini adalah proses *K-Means* pada atribut *Time In Hospital*. Dalam perhitungan ini menggunakan 10 data sampel dari *dataset* yang digunakan. Perhitungan ini menetapkan jumlah kelompok (*cluster*) berjumlah 3. Penentuan jumlah kelompok berdasarkan kebutuhan. Adapun langkah sebagai berikut:

1. Tentukan Nilai Centroid Awal (*Random*)

Tabel 3. Nilai Centroid Loop-1 Time In Hospital

Perulangan	Nama *Centroid*	Nilai *Centroid*
Loop - 1	*Centroid* 1	1
	Centroid 2	14
	Centroid 3	8

Pada tabel 3 merupakan penentuan nilai *centroid* awal atribut *time in hospital* secara *random*.

2. Perhitungan *Euclidean Distance* setiap *centroid* pada tabel 3.

Tabel 4. Nilai Euclidean Distance Centroid 1 Loop-1 Time In Hospital

No	Perhitungan	Hasil Euclidean Distance Centroid 1
1	$\sqrt{(1-1)^2}$	0
2	$\sqrt{(3-1)^2}$	2
3	$\sqrt{(14-1)^2}$	13
4	$\sqrt{(2-1)^2}$	1
5	$\sqrt{(14-1)^2}$	13
6	$\sqrt{(9-1)^2}$	8
7	$\sqrt{(13-1)^2}$	12
8	$\sqrt{(5-1)^2}$	4
9	$\sqrt{(8-1)^2}$	7
10	$\sqrt{(14-1)^2}$	13

Pada tabel 4 merupakan hasil perhitungan *Euclidean Distance Centroid 1 Loop-1 Time In Hospital* yang dimana nilai yang digunakan pada perhitungan ini berdasarkan nilai dari *dataset* atribut time in hospital dan nilai *centroid 1* pada tabel 3 dan menggunakan rumus 1. Untuk kolom hasil Euclidean Distance *Centroid 1* merupakan hasil dari perhitungan dari kolom perhitungan.

Untuk perhitungan *Euclidean Distance Centroid 2*, dan *Euclidean Distance Centroid 3* dilakukan seperti pada tabel 4 namun berdasarkan nilai *centroid* masing-masing pada tabel 3.

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

3. Menentukan Pengelompokan Berdasarkan Hasil *Euclidean Distance*

Tabel 5. Penentuan *Cluster Loop–1 Time In Hospital*

Loop – 1 Time In Hospital					
ED-1*	ED-2*	ED-3*	*Cluster 1*	*Cluster 2*	*Cluster 3*
0	13	7	1	-	-
2	11	5	1	-	-
13	0	6	-	1	-
1	12	6	1	-	-
13	0	6	-	1	-
8	5	1	-	-	1
12	1	5	-	1	-
4	9	3	-	-	1
7	6	0	-	-	1
13	0	6	-	1	-

**Note* :
- ED adalah *Euclidean Distance*. Dan nilai *Euclidean Distance* ini merupakan nilai yang telah dihitung sebelumnya
- Angka 1 pada *Cluster 1*, *Cluster 2*, Cluster 3 merupakan tanda bahwa data dikelompokkan pada *cluster*.

Dalam penentuan *cluster* dilakukan berdasarkan nilai dari *Euclidean Distance* yang telah di hitung sebelumnya pada tabel 4. Untuk penentuannya, dari tiap baris tersebut yang memiliki nilai *Euclidean Distance* terendah maka akan menjadi *cluster* dari kelompok *Euclidean Distance* tersebut. Adapun contoh pada tabel 5 untuk *Cluster Loop–1 Time In Hospital*.

4. Menetukan Pengelompokan Nilai Atribut *Time In Hospital*

Tabel 6. Hasil *Cluster Loop–1 Time In Hospital*

No.	*time_in_hospital*	*Cluster Loop–1*
1	1	*Cluster 1*
2	3	*Cluster 1*
3	14	*Cluster 2*
4	2	*Cluster 1*
5	14	*Cluster 2*
6	9	*Cluster 3*
7	13	*Cluster 2*
8	5	*Cluster 3*
9	8	*Cluster 3*
10	14	*Cluster 2*

Pada tabel 6 merupakan hasil dari penentuan *cluster loop–1 time in hospital* pada tabel 5. Dan nilai pada kolom *time in hospital* berdasarkan dari tabel *dataset* yang digunakan atribut *time in hospital*.

5. Menentukan Nilai *Centroid* Baru

Tabel 7. Nilai *Centroid Loop-2 Time In Hospital*

Perulangan	Nama *Centroid*	Perhitungan	Nilai *Centroid*
Loop – 2	*Centroid* 1	$(1 + 3 + 2)/3$	2
	Centroid 2	$(14 + 14 + 13 + 14)/4$	13.75
	Centroid 3	$(9 + 5 + 8)/3$	7.333333333

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

Pada tabel 7 merupakan penentuan nilai *centroid loop-2 atribut time in hospital* yang dimana nilai perhitungan berdasarkan hasil pengelompokan dari tabel 6 dan menggunakan rumus 2. Kolom nilai *centroid* merupakan hasil dari perhitungan dari kolom perhitungan.

Jika Nilai setiap *centroid* baru sama dengan nilai setiap *centroid* sebelumnya, maka proses *K-Means* selesai. Jika tidak ulangi proses dari langkah ke 2 yaitu perhitungan *Euclidean Distance* setiap *centroid* menggunakan nilai *centroid* baru.

Langkah 1-5 diatas juga dilakukan untuk menentukan pengelompokan pada atribut *num lab procedures, num procedures,* dan *num diagnoses*. Setelah dikelompokan pada proses *K-Means*. Hasil pengelompokan akan digunakan untuk proses *C4.5*.

B. Proses *C4.5*

Pada proses *C4.5* pertama tentukan atribut mana yang akan diprediksi. Dalam kasus ini atribut yang akan diprediksi adalah *A1cResult*. Untuk atribut yang akan diprediksi dilakukan perhitungan *entropy* pada tabel 8.

Tabel 8. Nilai *Entropy A1cResult*

Jenis Nilai	Jumlah	Perhitungan *Entropy*	*Entropy*
Norm	4	$\left(\frac{4}{10}\times(-1)\right)\times\log_2\left(\frac{4}{10}\right)$	0.528771238
>7	3	$\left(\frac{3}{10}\times(-1)\right)\times\log_2\left(\frac{3}{10}\right)$	0.521089678
>8	3	$\left(\frac{3}{10}\times(-1)\right)\times\log_2\left(\frac{3}{10}\right)$	0.521089678
Total Data	10	Total *Entropy (A1cResult)*	1.570950594

Setelah didapatkan nilai entropy *A1cResult* berdasarkan rumus 3, maka tahap selanjutnya hitunglah nilai *entropy* dan *gain* tiap atribut lainnya. Dalam perhitungan nilai gain tiap atribut lainnya akan bersangkutan dengan nilai total *entropy A1cResult*, yang dimana nilai *entropy A1cResult* akan digunakan dalam hitung gain total tiap atribut. Adapun contoh perhitungan *entropy* pada atribut *gender*.

Tabel 9. Nilai *Entropy Gender*

Jenis Nilai		Jumlah	Perhitungan *Entropy*	*Entropy*
Female	*Norm*	1	$\left(\frac{1}{5}\times(-1)\right)\times\log_2\left(\frac{1}{5}\right)$	0.464385619
	>7	2	$\left(\frac{2}{5}\times(-1)\right)\times\log_2\left(\frac{2}{5}\right)$	0.528771238
	>8	2	$\left(\frac{2}{5}\times(-1)\right)\times\log_2\left(\frac{2}{5}\right)$	0.528771238
Total Data *(Female)*		5	Total *Entropy (Female)*	1.521928095
Male	*Norm*	3	$\left(\frac{3}{5}\times(-1)\right)\times\log_2\left(\frac{3}{5}\right)$	0.442179356
	>7	1	$\left(\frac{1}{5}\times(-1)\right)\times\log_2\left(\frac{1}{5}\right)$	0.464385619
	>8	1	$\left(\frac{1}{5}\times(-1)\right)\times\log_2\left(\frac{1}{5}\right)$	0.464385619
Total Data *(Male)*		5	Total *Entropy (Male)*	1.370950594

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

Pada tabel 9 dilakukan perhitungan nilai *entropy* pada tiap jenis nilai yang terdapat pada atribut *gender*. Pada kolom jumlah, merupakan jumlah data yang dikelompokkan berdasarkan tiap jenis atribut *gender* yang memiliki nilai *norm*, >7, dan >8 pada atribut *A1cresult*. Perhitungan ini berdasarkan dari data hasil *K-Means* dan menggunakan rumus 3. Hasil nilai pada kolom *entropy* berdasarkan perhitungan dari kolom perhitungan *entropy*. Dan total *entropy* merupakan total dari nilai *entropy* tiap jenis nilai pada atribut *gender*.

Tabel 10. Nilai *Gain* Atribut *Gender*

Perhitungan	Total *Gain Gender*
$1.570950594 - \left(\left(\frac{5}{10} \times 1.521928095\right) + \left(\frac{5}{10} \times 1.370950594\right)\right)$	0.12451125

Pada tabel 10 merupakan perhitungan dan hasil nilai *gain* pada atribut *gender*. Perhitungan yang dilakukan pada tabel 10 menggunakan nilai total *entropy* atribut *A1cresult* pada tabel 8 dan menggunakan total data dan total *entropy* tiap jenis nilai pada atribut gender pada tabel 9 dan menggunakan rumus 4. Kolom total gain *gender* merupakan hasil dari perhitungan pada kolom perhitungan tabel 10.

Setelah dilakukan perhitungan *entropy* dan gain pada setiap atribut. Dari nilai gain semua atribut kecuali *A1cresult* carilah nilai *gain* yang tertinggi. Nilai gain tertinggi akan menjadi *root* pada pohon keputusan (*Decision Tree*). Pada kasus ini nilai *gain* tertinggi adalah atribut Diagnosis 1.

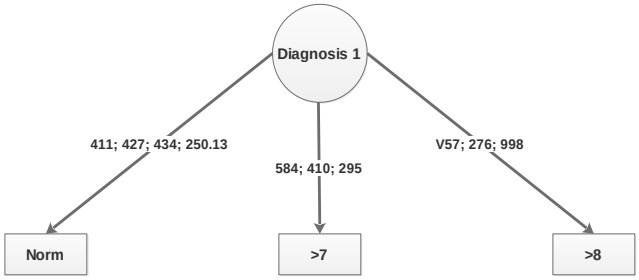
Tabel 11. *Root Decision*

Nama Atribut	Jenis Nilai				Jumlah
	Diagnosis 1	*Norm*	>7	>8	
Diagnosis 1	411	1	0	0	1
	427	1	0	0	1
	434	1	0	0	1
	250.13	1	0	0	1
	V57	0	0	1	1
	276	0	0	1	1
	998	0	0	1	1
	584	0	1	0	1
	410	0	1	0	1
	295	0	1	0	1

Pada tabel 11, bagian jenis nilai Diagnosis 1 digunakan untuk menjadi *leaf tree*. Untuk hasil keputusan tiap *leaf tree* berdasarkan jumlah jenis nilai *Norm*, >7, dan >8 dari tiap jenis nilai Diagnosis 1. Jika tiap jenis nilai Diagnosis 1 hanya memiliki satu jenis jumlah dari nilai *Norm*, >7, dan >8. Maka hasil keputusan pada *leaf tree* akan mengikuti nilai tersebut. Contoh jenis nilai Diagnosis 1 yaitu 411 hanya memiliki jumlah jenis nilai pada nilai *Norm*, sedangkan >7 dan >8 berjumlah 0. Maka hasil keputusan pada nilai 411 adalah *Norm*. Jika tiap jenis nilai atribut memiliki jumlah nilai selain salah satu dari *Norm*, >7, dan >8. Maka hasil pada *leaf tree* pada jenis nilai atribut tersebut harus ditentukan dengan mengulangi kembali semua langkah dari awal C4.5 kecuali atribut yang telah menjadi *tree* tidak perlu dihitung kembali. Pada *modeling* ini, hasil dari tiap nilai jenis atribut Diagnosis 1 menunjukkan hanya memiliki 1 nilai dari 3 jenis nilai *A1cResult*.

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

Sehingga perhitungan selesai dan tidak perlu melakukan perulangan. Adapun hasil decision tree sebagai berikut (gambar 7):



Gambar 7. *Decision Tree C4.5*

2.5. *Evaluation*

Mengevaluasi model dan hasil yang dilakukan pada tahap modeling. Adapun hasilnya dalam kasus ini adalah jumlah data yang digunakan akan mempengaruhi mempengaruhi bentuk *decision tree*. Dalam kasus ini menggunakan 10 data sebagai *sample*.

2.6. Deployment

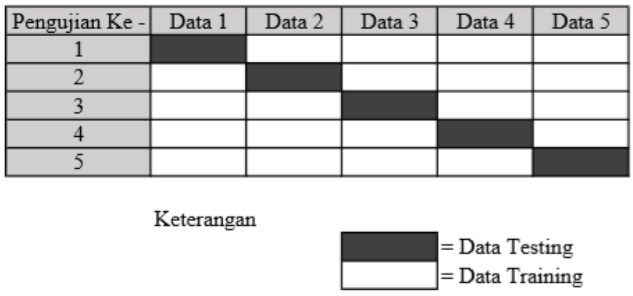
Menerapkan hasil dari model yang didapatkan dari tahapan modeling. Adapun implementasi diterapkan pada pembuatan aplikasi untuk memprediksi *HbA1c* dan didalam aplikasi tersebut terdapat pengimplementasian metode *K-Means* dan *C4.5*.

2.7. *Validasi Akurasi*

Validasi akurasi bukan merupakan tahapan *CRISP-DM*. Tahapan ini akan menguji validasi tingkat keakurasian hasil prediksi HbA1c terhadap metode *K-Means* dan *C4.5*. Dalam pengujian validasi akurasi menggunakan *K-Fold Cross Validation*. Adapun rumus *K-Fold Cross Validation* sebagai berikut rumus (5) [11]:

$$\text{akurasi} = \frac{\sum \text{klasifikasi benar}}{\sum \text{data uji}} \times 100\% \qquad (5)$$

Keterangan:
akurasi = hasil akurasi
klasifikasi benar = jumlah prediksi benar
data uji = jumlah data yang dilakukan pengujian



Gambar 8. Pembagian Data *Training* dan Data *Testing*

Pada gambar 6 merupakan pembagian data *training* dan data *testing* pada tiap skema pengujian. Yang dimana pada kasus ini setiap kelompok data terdapat 5 jumlah data didalamnya.

PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika
Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

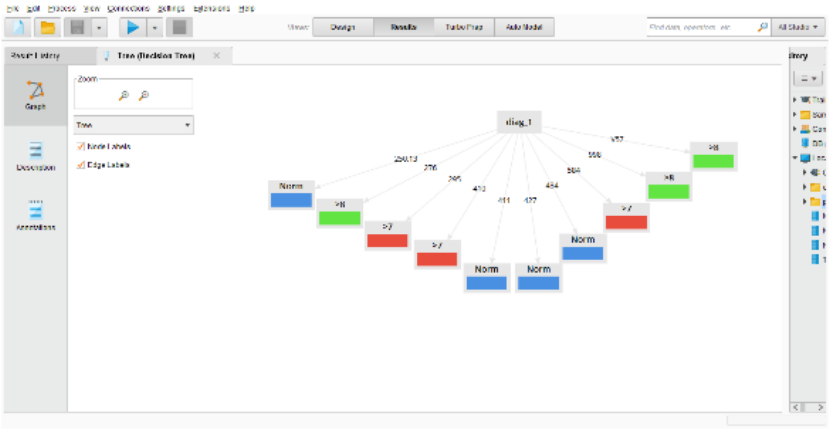
Tabel 12. Perhitungan Akurasi *K-Fold Cross Validation*

Pengujian Ke -	Jumlah Prediksi Benar	Jumlah Prediksi Salah	Perhitungan	Akurasi
1	2	3	$\frac{(2+4+4+4+4)}{25}(\times 100\%)$	72%
2	4	1		
3	4	1		
4	4	1		
5	4	1		

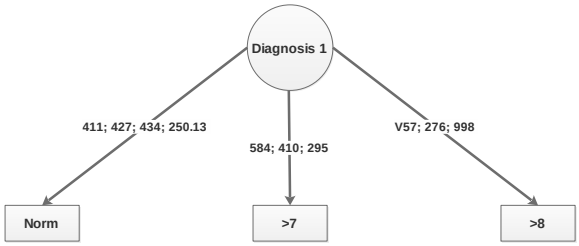
Tabel 12 merupakan perhitungan akurasi *K-Fold Cross Validation* yang dimana nilai pada kolom jumlah prediksi benar dan jumlah prediksi salah berdasarkan data pada skema gambar 5. Dan pada tabel 12 menggunakan rumus 5. Kolom akurasi pada tabel 12 merupakan hasil perhitungan dari kolom perhitungan.

3. HASIL DAN PEMBAHASAN

3.1. Perbandingan Hasil

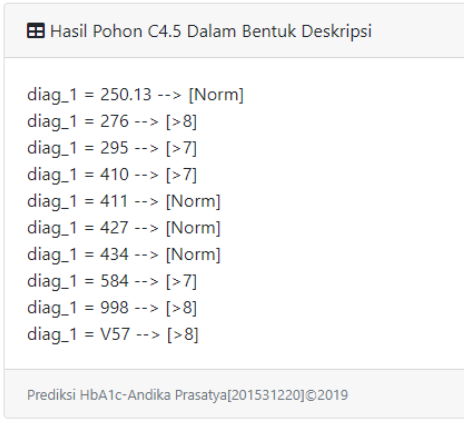


Gambar 9. Rapidminer Hasil Decision Tree C4.5



Gambar 10. *Microsoft Excel* Hasil *Decision Tree C4.5*

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925



Gambar 11. Aplikasi Hasil *Decision Tree C4.5*

Dari hasil pada gambar 7, 8, dan 9. Didapatkan kesimpulan bahwa hasil *decision tree C4.5* pada *rapidminer*, *microsoft excel*, dan aplikasi adalah memiliki hasil yang sama. Pada hasil *decision tree C4.5* didapatkan bahwa atribut yang menjadi *root* adalah atribut diagnosis 1. Hasil yang didapatkan dari setiap *decision tree* pada *rapidminer*, *microsoft excel*, dan aplikasi yang proses dari *K-Means* hingga *C4.5* berdasarkan dari tiap-tiap *tools* tersebut.

3.2. Implikasi

Dari hasil penelitian yang didapatkan bahwa data pasien penderita diabetes pada penelitian ini dapat digunakan pada metode *K-Means* dan *C4.5*. Dikarenakan dari penelitian dengan menggunakan metode *K-Means* dan *C4.5* pada data penderita diabetes didapatkan hasil prediksi dan akurasi. Akurasi hasil validasi menggunakan *K-Fold Cross Validation* sebesar 72%. Berdasarkan hasil akurasi yang didapatkan tersebut, maka prediksi *HbA1c* dapat digunakan sebagai salah satu solusi untuk mengatasi keterbatasan laboratorium dalam memberikan pelayanan pemeriksaan *HbA1c* pada penderita diabetes.

4. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Hasil penelitian ini dapat digunakan sebagai salah satu solusi untuk mengatasi keterbatasan pada laboratorium dalam memberikan alternatif keputusan pelayanan pemeriksaan HbA1c pada penderita diabetes yang akan melakukan kontrol / konsultasi ke dokter. Namun adapun ketentuan atau syarat supaya hasil penelitian ini dapat dijadikan solusi yaitu dalam proses *training* pada metode *K-Means* dan *C4.5* yang digunakan untuk menentukan pola atau *rule* prediksi HbA1c diharuskan menggunakan data *training* yang banyak, bervariasi, dan terdapat unsur unik pada data tersebut. Jika tidak memenuhi ketentuan tersebut maka yang terjadi ketika ingin memprediksi *HbA1c* pada penderita diabetes yang dimana data penderita tersebut nilainya tidak ada dalam pola atau *rule* maka hasil prediksi *HbA1c* yang didapatkan memiliki tingkat keakurasian yang rendah dan bahkan juga hasil yang didapatkan adalah tidak diketahui.

Dan juga hasil akurasi validasi yang didapatkan dengan *K-Fold Cross Validation* terdapat pada tabel 12 sebesar 72%

4.2. Saran

Pengembangan kedepannya dapat menggunakan metode yang berbeda dari metode pada penelitian ini sehingga dapat dibandingkan. Dari perbandingan tersebut akan menghasilkan kesimpulan metode mana yang menghasilkan akurasi prediksi *HbA1c* yang terbaik.

Vol. 13, No. 1, Maret 2020, P-ISSN 1978-9262, E-ISSN 2655-5018
DOI: https://doi.org/10.33322/petir.v13i1.925

DAFTAR PUSTAKA

[1] C. Buell, D. Kermah, and M. B. Davidson, "Utility of A1C for Diabetes Screening in the 1999 2004 NHANES Population," Diabetes Care, vol. 30, no. 9, pp. 2233–2235, Sep. 2007.

[2] S. R. Paputungan and H. Sanusi, "Peranan Pemeriksaan Hemoglobin A1c pada Pengelolaan Diabetes Melitus," Cdk-220, vol. 41, no. 9, pp. 650–655, 2014.

[3] S. Bahri, D. M. Midyanti, and R. Hidayati, "Perbandingan Algoritma Naive Bayes dan C4 . 5 Untuk Klasifikasi Penyakit Anak," Semin. Nas. Apl. Teknol. Inf., vol. 07, no. 22, pp. 24–31, 2018.

[4] L. Zahrotun, "Analisis Pengelompokan Jumlah Penumpang Bus Trans Jogja Menggunakan Metode Clustering K-Means Dan Agglomerative Hierarchical Clustering (AHC)," J. Inform., vol. 9, no. 1, pp. 1039–1047, Jan. 2015.

[5] V. Derbentsev, N. Datsenko, O. Stepanenko, and V. Bezkorovainyi, "Forecasting Cryptocurrency Prices Time Series Using Machine Learning Approach," SHS Web Conf., vol. 65, p. 02001, May 2019.

[6] S. A. Soelistijo et al., Konsensus Pengelolaan Dan Pencegahan Diabetes Melitus Tipe 2 Di Indonesia 2015, no. 11. PB. PERKENI, 2015.

[7] B. Strack et al., "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," Biomed Res. Int., vol. 2014, 2014.

[8] M. Nishom and D. S. Wibowo, "Implementasi Metode K-Means berbasis Chi-Square pada Sistem Pendukung Keputusan untuk Identifikasi Disparitas Kebutuhan Guru," J. Sist. Inf. Bisnis, vol. 8, no. 2, p. 187, Nov. 2018.

[9] M. Mirqotussa'adah, M. A. Muslim, E. Sugiharti, B. Prasetiyo, and S. Alimah, "Penerapan Dizcretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4.5 dalam Mendiagnosa Diabetes," Lontar Komput. J. Ilm. Teknol. Inf., no. September, p. 135, Aug. 2017.

[10] A. Paramitha Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," JuTISI, vol. 1, pp. 260–270, Apr. 2015.

[11] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 5, p. 577, Oct. 2018.

[12] Budiana ND, Siregar RR, Susanti MN. Penetapan Instruktur Diklat Menggunakan Metode Clustering K-Means dan Topsis Pada PT PLN (Persero) Udiklat Jakarta. PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika. 2019 Aug 6;12(2):111-21.

[13] Siregar RR, Nasution FS. Algoritma C4. 5 Untuk Prediksi Tingkat Kelulusan Mahasiswa Jurusan Teknik Informatika STT PLN. Jurnal Informatika dan Komputasi. 2017 Mar 1;9(1):1-6.

[14] Setiyadi D. Data Mining Dengan Metode Decision Tree Algoritma C4. 5 Untuk Memprediksi Permintaan Jenis Produk Barang. Inesit 2019 Apr 30 (Vol. 6, No. 2, Pp. 13-34).

[15] Patel HG, Sarvakar K. Research Challenges and Comparative Study of Various Classification Technique Using Data Mining. International Journal of Latest Technology in Engineering, Management & Applied Science. 2014;3(9):170-6.