2st SEMINAR NASIONAL TEKNOLOGI DAN MULTIDISIPLIN ILMU
SEMNASTEKMU 2022
Universitas Sains dan Teknologi Komputer Semarang, 11 Desember 2022
Available online at**:** **https://prosiding.stekom.ac.id/index.php/semnastekmu**
E-ISSN: 2809-1566
P-ISSN: 2809-1574

KOMPARASI UJI PERFORMA ALGORITMA C4.5 DAN *K-NEAREST NEIGBOR* DALAM MEMPREDIKSI PENYAKIT DIABETES

Adhi Tristiyanto [1], Sriyanto [2]
[1,2] Program Studi Magister Teknik Informatika, Fakultas Ilmu Komputer,
e-mail: [1]adhi.tristiyanto.2021211001@mail.darmajaya.ac.id, [2]sriyanto@darmajaya.ac.id
Institut Informatika Dan Bisnis Darmajaya

Alamat: Jl. ZA. Pagar Alam No.93, Gedong Meneng, Kec. Rajabasa, Kota Bandar Lampung, Lampung 35141;Telepon: (0721) 787214
Korespondensi penulis : adhi.tristiyanto.2021211001@mail.darmajaya.ac.id

ABSTRACT

Tingginya angka statistik penderita diabetes memerlukan antisipasi dari dinas kesehatan untuk menekan dan mencegah ledakan penderita diabetes. Oleh karena itu, diagnosis dini diabetes dapat membantu pengobatan diabetes lebih cepat dan dapat menghindari komplikasi penyakit berbahaya lainnya. Salah satu pencatatan yang dapat dilakukan adalah dengan memanfaatkan teknik klasifikasi dengan data mining. Algoritma C4.5 dan K-Nearest Neigbor merupakan metode yang dapat digunakan untuk memprediksi diabetes. Dari hasil pengujian 520 data yang terbagi menjadi 80% atau 416 data sebagai data latih dan 20% atau 104 sebagai data pengujian, algoritma C45 (Decision Tree) memiliki akurasi yang lebih tinggi dibandingkan k-NN dengan menggunakan seluruh atribut dengan akurasi sebesar 98% sedangkan k-NN memiliki akurasi sebesar 83%. Sedangkan untuk penggunaan menggunakan fitur berbasis heatmap, algoritma C45 memiliki akurasi 88% lebih tinggi dibandingkan algoritma k-NN dengan akurasi 84%. Pada algoritma k-NN, mencari jumlah k yang optimal adalah 4 dengan akurasi 88%.

Kata Kunci: Algoritma C4.5, K-Nearest Neighbor, Diabetes Mellitus

ABSTRACT

The high statistical number of diabetics requires anticipation from the health department to suppress and prevent an explosion in diabetes sufferers. Therefore, early diagnosis of diabetes can help treat diabetes more quickly and can avoid complications of other dangerous diseases. One of the records that can be done is to utilize classification techniques with data mining. The C4.5 and K-Nearest Neigbor algorithms are methods that can be used to predict diabetes. From the test results of 520 data which is divided into 80% or 416 data as training data and 20% or 104 as test data, the C45 (Decision Tree) algorithm has higher accuracy than k-NN by using all attributes with an accuracy of 98% while k-NN has an accuracy of 83%. As for using heatmap-based features, the C45 algorithm has an accuracy of 88% higher than the k-NN algorithm with an accuracy of 84%. In the k-NN algorithm, finding the optimal number of k is 4 with an accuracy of 88%.

Keywords: *Algorithm C4.5, K-Nearest Neighbor, Diabetes Mellitus*

1. PENDAHULUAN

Diabetes tidak hanya menyebabkan kematian prematir di suluruh dunia tetap juga menjadi penyebab utama kebutaan, penyakit jantung dan gagal ginjal [1]. Menurut International Diabetes Federation (IDF) memperkirakan data penderita diabetes berdasarkan jenis kelamin menyebutkan prevelensi diabetes pada tahun 2019 mencapai 9% pada perempuan dan 9,65% pada laki-laki dan Indonesia menjadi satu-satunya Negara di Asia Tenggara dengan jumlah penderita tertinggi [1]. Melihat tingginya angka statistik penderita penyakit diabetes maka patut adanya antisipasi oleh pihak layanan kesehatan untuk menekan dan mencegah timbulnya ledakan pasien diabetes. Salah satu yang pencatatan yang bisa dilakukan adalah dengan memanfaatkan teknik klasifikasi dengan data mining[2] [2]. Klasifikasi banyak digunakan untuk menentukan keputusan sesuai pengetahuan baru yang didapat dari pengolahan data lampau menggunakan perhitungan suatu algoritma [3].

Pada penelitian sebelumnya kasus mengenai klasifikasi diabetes telah banyak dilakukan dengan memanfaatkan teknologi informasi untuk membantu klasifikasi penyakit diabetes dengan menggunakan beberapa metode diantaranya Agung Mulyo Widodo dkk (2021) yang menguji performa KNN, Naïve Bayes dan Regresi Logistik untuk mengklasifikasi diabetes dengan diperoleh hasil diperoleh bahwa algoritma K-NN menghasilkan model prediksi dengan akurasi yang tertinggi dibandingkan ketiga algoritma yang digunakan dalam penelitian ini. Sementara Maulidya Dwi Nurmalasari, dkk (2021) mengkomparasi algoritma Naïve Bayes dan KNN untuk membangun pengetahuan diagnose Penyakit Diabetes dengan hasil algoritma *K- Nearest Neighbor* sebagai algoritma yang memiliki akurasi terbaik.

Sementara Marcos dan Utomo (2015) melakukan perbandingan kinerja lagoritma C4.5 dan Naïve Bayes untuk mengklasifikasi penyakit diabetes dan diperoleh hasil akhir diperoleh bahwa algoritme naïve bayes lebih baik dari pada algoritme C4.5 karena memiliki tingkat akurasi yang lebih baik. Penelitian yang dilakukan oleh Ardiansyah dkk (2021) dengan melakukan analisis perbandingan akurasi algoritma naïve bayes dan C4.5 untuk klasifikasi Diabetes diperoleh hasil bahwa algoritma C4.5 memiliki hasil yang baik dalam klasifikasi penyakit diabetes dibandingkan algoritma Naïve Bayes dengan tingkat accuracy 99.03%, precision 100%, dan recall 98.18%. Berdasarkan uraian latar belakang dan penelitian sebelumnya maka peneliti menggunakan data mining dengan model C4.5 Dan K-Nearest Neigbor untuk mengklasifikasi penderita penyakit diabetes.

2. TINJAUAN STUDI

Data mining adalah proses menemukan pola dan tren yang berguna dalam kumpulan data yang besar[4]. Hasil data mining sebagai langkah penting untuk menemukan pengetahuan dalam proses database dengan mengekstrak pengetahun tersembunyi seperti pola, hubungan atau aturan dari kumpulan data besar sehingga dapat dianalisis dan mampu digunakan untuk memprediksi tren masa depan [5]. Teknik data mining berdasarkan tugas yang dapat dilakukan untuk tujuan prediksi dan deskripsi antara lain: *Classification, Regression, Clustering, Summarization, Dependency modeling dan Change and deviation detection* [6]. Tugas data mining yang paling umum dapat ditemukan di hampir setiap bidang usaha seperti perbankan, pendidikan, kedokteran, hukum dan keamanan adalah klasifikasi [4]. Pengklasifikasi menghasilkan model klasifikasi berdasarkan data pelatihan berisi objek yang dijelaskan oleh nilai yang mereka miliki pada sekumpulan atribut, satu atribut dibedakan sebagai kelas (Asif,dkk., 2017).

Algoritma k-nearest neighbor adalah algoritma yang paling sering digunakan untuk klasifikasi [7]. *K-Nearest Neighbor* (K-NN) termasuk kelompok *instance-based learning dan* dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing [8]. Algoritma metode k-nearest neighbor (KNN) bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan KNN-nya

[9]. kNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing [10].

Algoritma C4.5 merupakan kelompok algoritma *Decision Tree* [11]. Decision tree merupakan pohon keputusan classifier yang mengklasifikasikan data ke dalam label kelas yang telah ditentukan [12]. Inti dari pohon keputusan mencakup simpul akar tunggal, beberapa simpul internal dan beberapa simpul daun dan Setiap simpul daun memegang label kelas kemudian Jalur dari simpul akar ke simpul daun mengungkapkan aturan klasifikasi [13].

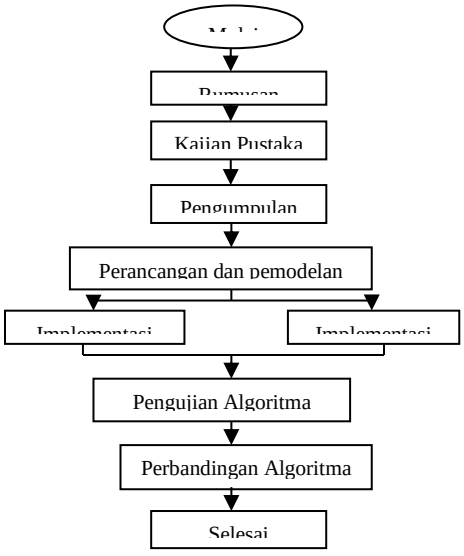
Beberapa penelitian terkait yang telah menguji penerapan data mining pada beberapa objek seperti Agung Mulyo Widodo dkk (2021) yang menguji Performa dari K-NN, J48, Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes dan dihasilkan algoritma K-NN menghasilkan model prediksi dengan akurasi yang tertinggi dibandingkan ketiga algoritma, Maulidya Dwi Nurmalasari , dkk (2021) yang menguji Algoritma Naïve Bayes dan K-Nearest Neighbor untuk Membangun Pengetahuan Diagnosa Penyakit Diabetes dan hasil menunjukkan bahwa terbukti algoritma *K- Nearest Neighbor* sebagai algoritma yang memiliki akurasi terbaik, Mursyid Ardiansyah , dkk (2021) menguji Akurasi Algoritma Naïve Bayes dan C4.5 untuk Klasifikasi Diabetes dan hasil menunjukkan bahwa algoritma C4.5 (skenario 4) memiliki hasil yang baik dalam klasifikasi penyakit diabetes, Wisti Dwi Septiani dan Untung Rohwadi (2021), Rini Andanika Siallagan dan Fitriyani (2021), Fida Maisa Hana (2020) menguji Algoritma Algoritma C4.5 Untuk Klasifikasi Penderita Penyakit Diabetes dan diperoleh hasil akurasi di atas 90%.

Berdasarkan hasil penelitian terkait yang telah disebutkan pada pada tabel 2.2 diatas maka menjadi dasar atau alasan penulis untuk menggunakan teknik data mining metode algoritma C4.5 dan k-NN untuk melakukan penelitian di bidang data mining bidang kesehatan untuk mengklasifikasi penderita diabetes.

3. METODE PENELITIAN

3.1 Objek dan MetodePenelitian

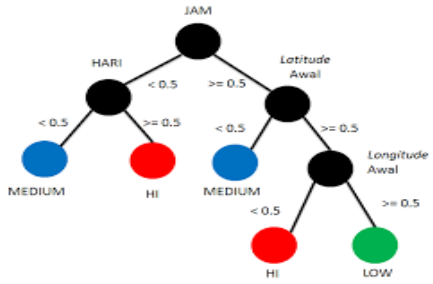
Penelitian dilakukan pada penderita penyakit diabetes yang diperoleh dari kaggle.com/datasets dengan nama file diabetes. Metode penelitian dalam mengembangkan data mining mengikuti beberapa tahapan yang dapat dilihat pada gambar 2.1 bawah ini:



Gambar 3.1 Tahapan alur penelitian

3.2 Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (*Decision Tree*). Pohon keputusan berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah variabel input dengan sebuah variabel target.



Gambar 3.2 Ilustrasi Decision Tree

Langkah-langkah untuk membangun pohon keputusan adalah sebagai berikut [12] :
- Pilih sebuah atribut sebagai *root*.
- Membuat cabang untuk setiap nilai.
- Perhatikan kasus di cabang.
- Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama

Nilai gain tertinggi dari atribut-atribut yang ada digunakan sebagai dasar untuk memilih atribut sebagai akar. Untuk mendapatkan nilai gain, terlebih dahulu menghitung nilai entropy atribut data dengan persamaan sebagai berikut:

$$\mathbf{Entropy(S)} = -\sum_{i=1}^{n} \mathbf{p_i} * \mathbf{\log_2 p_i}$$

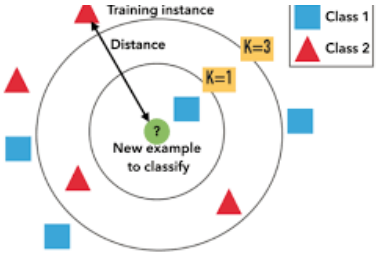
Keterangan :
S = Himpunan kasus
n = Jumlah partisi S
pi = Proporsi dari Si terhadap S

Akar dari pohon keputusan algoritma C4.5 adalah atribut data dengan nilai gain tertinggi. Adapun formula untuk menghitung nilai gain dapat dilihat pada persamaan sebagai berikut:

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

3.3 Algoritma k-Nearest Neighbor (k-NN)

Algoritma k-Nearest Neighbor adalah algoritma *supervised learning* dimana hasil dari *instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori k-tetangga terdekat.Tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sample-sample dari *training* data.



Gambar 3.3 Ilustrasi K Nearest Neighbor

Langkah yang digunakan dalam metode K-Nearest Neighbor :
- Tentukan parameter K (jumlah tetangga paling dekat).
- Hitung kuadrat jarak euclid masing masing objek terhadap data sample yang diberikan.
- Urutkan objek objek kedalam kelompok yang memiliki jarak terkecil.
- Kumpulkan kategori Y (Klasifikasi nearest neighbor).
- Dengan kategori nearest neighbor yang paling banyak, maka dapat diprediksikan nilai query instance yang telah dihitung.
- Perhitungan jarak terdekat menggunakan algoritma euclidien seperti yang ditunjukkan pada persamaan sebagai berikut (Argina, 2020):

$$euc = \sqrt{((a\_1 - b\_1)^2 + \cdots + (a\_n - b\_n)^2)}$$

Keterangan:
a = a1,a2…,an, b = b1,b2....,bn dimana mewakili n nilai atribut dari dua record. Untuk atribut dengan nilai kategori
Perhitungan jarak yang paling umum dipakai pada perhitungan pada algoritma KNN adalah menggunakan perhitungan jarak Euclidean. Rumusannya adalah sebagai berikut:

$$euc = \sqrt{(\sum_{i=1}^{n}(pi - qi)^2)}$$

Keterangan :
pi = sample data
qi = data uji
i = variabel data
n = dimensi data

3.4 Pengujian Algoritma

Langkah awal dalam tahap ini adalah dengan membagi data di setiap kasus menjadi 2, yaitu data training atau data latih dan data testing atau data uji. Data training digunakan sebagai data rujukan dalam perhitungan setiap algoritma, sedangkan data testing digunakan untuk menilai prediksi maupun penentuan yang dilakukan oleh setiap algoritma sudah tepat atau tidak. Pengujiann dilakukan dengan mengasosiasikan nilai prediksi dengan nilai sebenarnya. [14] menyebutkan ada empat kemungkinan hasil prediksi yaitu *true positive* (TP) dan juga sebagai *true negative* (TN) yang dapat dijelaskan pada tabel 2.1 dibawah ini:

Tabel 2.1. *Confusion Matrix*

	Predicted Negative	Predicted Positiv
Actual Negative	True Negative (TN)	False Positive (FN)
Actual Positive	False Negative (FN)	True Positive (TP)

Melalui empat kemungkinan prediksi diatas maka berbagai metrik kinerja dapat digunakan untuk mengukur kualitas hasil klasifikasi yaitu:
Accuracy merupakan Total keseluruhan seberapa sering model benar mengklasifikasi. Formula accuracy dapat ditulis sebagai berikut :

$$Accuracy = \frac{TP+TN}{Total} \text{..... (1)}$$

Pecision yaitu ketika model memprediksi positif maka seberapa sering prediski itu benar. Formula *Pecision* dapat ditulis sebagai berikut :

$$Precision = \frac{TP}{FP+TP} \text{ ..... (2)}$$

Recall yaitu Ketika kelas aktualnya positif, seberapa sering model memprediksi positif. Formula *Recall* dapat ditulis sebagai berikut :

$$Recall = \frac{TP}{FN+TP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3)}$$

3.5 Perbandingan algoritma

Setelah tahap pengujian model maka langkah berikutnya adalah melakukan perbandingan nilai *precision*, *recall* dan *accuracy* pada masing-masing algoritma di setiap kasus. Langkah selanjutnya dilakukan rekapitulasi hasil dari masing-masing algoritma sehingga dapat diambil kesimpulan mengenai algoritma terbaik.

3.6 Alat dan bahan

Alat dan *bahan* yang digunakan dalam penelitian ini antara lain perangkat keras berupa laptop dengan spesifikasi processor amd a10, hdd 100 gb, os linux, vga radeon r5+r6. Sementara untuk perangkat lunak yang dipakai adalah *google collab Chrome library* dan *scikit learn* merupakan *tools* yang dimanfaatkan untuk mengimplementasikan metode *data mining* yang digunakan, Algoritma C4.5 dan k-NN sebagai Algoritme perhitungan untuk menyelesaikan klasifikasi penderita penyakit diabetes dan teknik pengujian *confusion matrix* diterapkan untuk memvalidasi nilai akurasi model yang dibangun.

4. HASIL DAN PEMBAHASAN

4.1 Eksplorasi Data

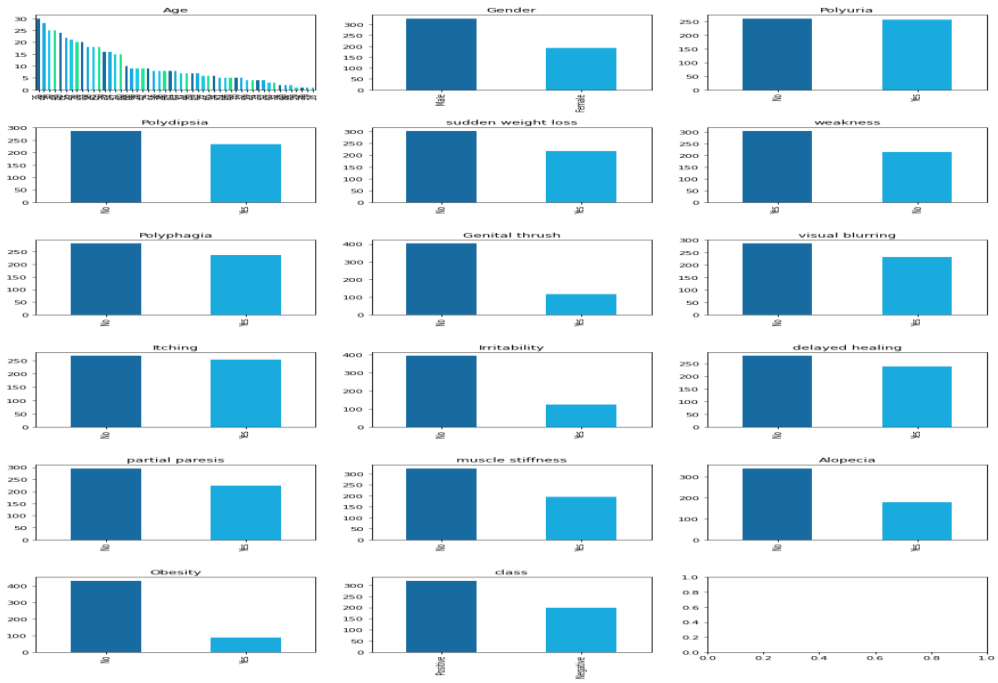
Dataset awal memiliki beberapa atribut dimana terdapat label dengan nama '*class'* dan sisanya adalah *feature*. Adapun tampilan dataset awal dapat dilihat pada Gambar 4.1 sebagai berikut:

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
58	Male	No	Yes	No	No	No	No	Yes	Yes	No	No	No	Yes	No	No	Negative
70	Male	Yes	No	No	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Negative
61	Male	No	No	No	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes	No	Negative
39	Female	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Positive
50	Female	Yes	Yes	Yes	No	Yes	No	No	No	No	Yes	Yes	No	No	No	Positive

Gambar 4.1 data dari dataset

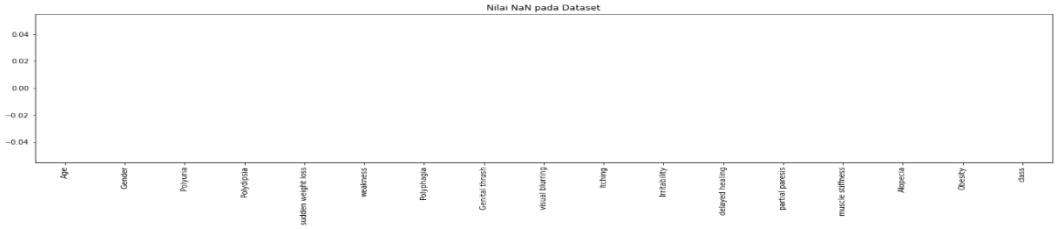
4.2 *Tahap preprocessing*

Berdasarkan dataset yang sudah disiapkan sebelumnya maka proses selanjutnya adalah melakukan visualisasi pada dataset untuk mengetahui *preprocessing* apa saja yang harus dilakukan sebagai tahapan pembersihan data. Berikut visualisasi dataset dapat dilihat pada gambar 4.2 sebagai berikut:



Gambar 4.2 Visualisasi dataset

Hasil dari visualisasi diatas, dapat disimpulkan bahwa hampir semua atribut dalam dataset merupakan data kategorikal dengan tipe *string*. Hanya pada atribut '*age*' yang memiliki nilai *integer*. Langkah selanjutnya adalah melakukan visualisasi terhadap nilai NaN atau nilai kosong pada kolom. Adapun hasil visualisasi nilai NaN dapat dilihat pada gambar 4.3 sebagai berikut:



Gambar 4.3 Visualisasi nilai NaN

Tahapan selanjutnya adalah mengubah nilai atribut menjadi bentuk numerik dengan menggantinya menjadi 0 dan 1.Berikut adalah dataset yang sudah melalui proses *preprocessing* dapat dilihat pada gambar 4.4 sebagai berikut:

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability
38	0	1	1	1	1	1	0	1	1	1
53	1	0	0	0	1	1	0	1	1	0
48	0	1	1	0	1	0	0	1	1	0
44	1	1	1	0	1	0	1	0	0	1
38	0	1	1	1	1	1	0	0	0	0

delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
1	1	1	0	0	1
1	1	1	1	0	0
1	1	0	0	0	1
1	0	1	1	0	1
0	1	0	0	0	1

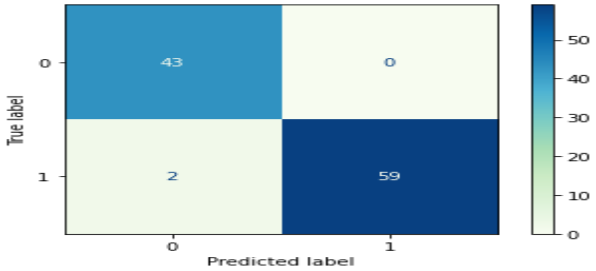
Gambar 4.4 Dataset setelah preprocessing

Setelah dilakukan *preprocessing*, selanjutnya dilakukan pemisahan antara data *train* yaitu data untuk proses pelatihan dan data *test* yaitu data untuk menguji hasil pelatihan yang dilakukan dengan menggunakan data *train*. Jumlah data dalam penelitian ini sebanyak 520 data yang kemudian data

tersebut dibagi menjadi data training dan data testing dengan presentase 80% data sebagai data *training* dan 20% sebagai data *testing*. Hasil perhitungan persentase dari masing-masing data diperoleh nilai 80% dari jumlah data 520 didapatkan data *training* sebanyak 416 data dan 20% dari jumlah data 520 didapatkan data *testing* sebanyak 104 data.

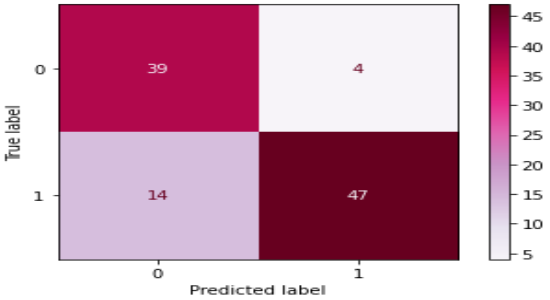
4.3 Evaluasi

Tahap evaluasi pada penelitian ini dilakukan dengan menggunakan seluruh atribut yang ada pada dataset dan didapatkan hasil dalam bentuk confusion matrix pada masing-masing algortima seperti pada gambar 4.5 sebagai berikut :



Gambar 4.5 Confusion matrix C45

Berdasarkan gambar 4.5 perhitungan confusion matrix C4.5 diperoleh nilai 43 true negative dimana model memprediksi 43 data tersebut berlabel negative dan pada data asli juga bertipe negative. Sementara pada 59 model memprediksi data berlabel positif dan pada data asli juga berlabel positif tetapi model memprediksi diperoleh dengan salah sebanyak 2 data. Perhitungan confusion matrix k-NN diperoleh hasil seperti pada gambar 4.6 sebagai berikut :



Gambar 4.6 Confusion matrix K Nearest Neighbor

Berdasarkan gambar 4.6 perhitungan confusion matrix k-NN diperoleh nilai 39 true negative dimana model memprediksi 39 data tersebut berlabel negative dan pada data asli juga bertipe negative. Sementara pada 47 model memprediksi data berlabel positif dan pada data asli juga berlabel positif tetapi terdapat 4 data yang salah atau type eror false positive. Dari perhitungan confusion matrix diperoleh juga hasil *classification report* dari kedua model dengan seperti yang terlihat pada gaambar 4.7 dan 4.8 sebagai berikut:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	43
1	1.00	0.97	0.98	61
accuracy			0.98	104
macro avg	0.98	0.98	0.98	104
weighted avg	0.98	0.98	0.98	104

Gambar 4.7 Classification report C45

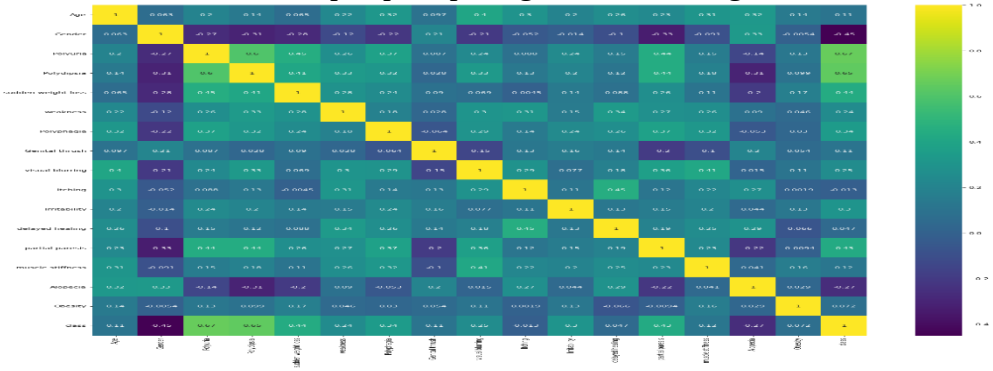
	precision	recall	f1-score	support
0	0.74	0.91	0.81	43
1	0.92	0.77	0.84	61
accuracy			0.83	104
macro avg	0.83	0.84	0.83	104
weighted avg	0.84	0.83	0.83	104

Gambar 4.8 Classification report K Nearest Neighbor

Dari *confusion matrix* dan *classification report* pada kedua algorima diatas maka dapat disimpulkan bahwa algoritma C45 lebih unggul dengan nilai akurasi 98% dan eror *false negative* hanya 2 data. Sedangkan dengan menggunakan algoritma *K-Nearest Neighbor* diperoleh nilai akurasi 83% dengan *false negative* 14 data dan *false positive* 4 data. Pengujian pada penelitian ini dilakukan dengan menggunakan data *test*.

4.4 Eksperimen

Pada tahap ini, peneliti mencoba melakukan eksperimen untuk menemukan akurasi maksimal dari dataset dengan menggunakan kedua algoritma yaitu C4.5 dan k-NN. Dilakukan *feature selection* berdasarkan hasil korelasi dari *heatmap* seperti pada gambar 4.9 sebagai berikut:



Gambar 4.9 Heatmap

Berdasarkan hasil korelasi dari *heatmap* didapatkan fitur yang memiliki korelasi positif dimana ketika label bernilai 1 maka fitur juga bernilai 1 dan korelasi negative. Kemudian ketika label bernilai 1 maka fitur juga bernilai 0. Fitur yang diambil adalah data yang lebih dari 0,3 dan kurang dari -0,3 yaitu fitur *'Gender', 'Polydipsia', 'sudden weight loss',* dan *'Polyphagia'*, didapatkan *classification report* dari kedua algoritma seperti pada gambar 4.10 dan gambar 4.11 sebagai berikut:

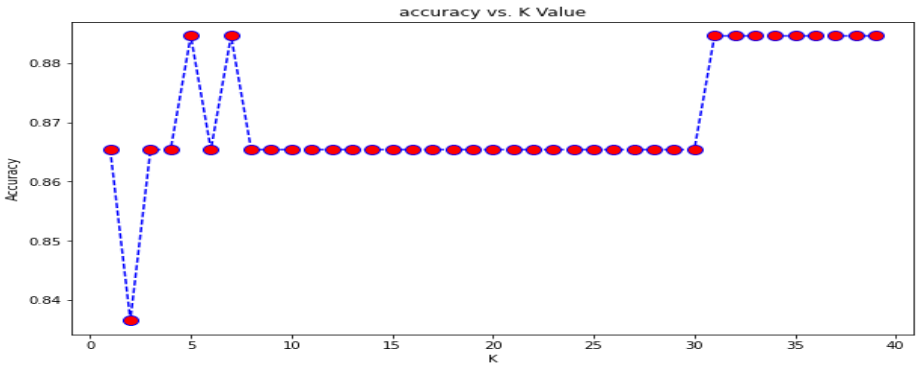
	precision	recall	f1-score	support
0	0.78	0.98	0.87	48
1	0.98	0.77	0.86	56
accuracy			0.87	104
macro avg	0.88	0.87	0.87	104
weighted avg	0.89	0.87	0.86	104

Gambar 4.10 Classification report C45

	precision	recall	f1-score	support
0	0.75	0.98	0.85	48
1	0.98	0.71	0.82	56
accuracy			0.84	104
macro avg	0.86	0.85	0.84	104
weighted avg	0.87	0.84	0.83	104

Gambar 4.11 Classification report K Nearest Neighbor

Berdasarkan *classification report* maka dapat dilakukan eksperimen untuk menemukan nilai k yang optimum untuk algoritma K Nearest Neighbor dengan dilakukan percobaan dari k 1 sampai 40 dengan hasil seperti pada gambar 4.12 sebagai berikut:



Gambar 4.12 Visualisasi percobaan pada k-NN

Berdasarkan hasil visualisasi percobaan pada k-NN, diperoleh nilai k yang optimum untuk algoritma KNN adalah 4 dengan akurasi 88%

4.5 Perbandingan Algoritma

Berdasarkan hasil yang didapatkan dari tahapan awal yaitu eksplorasi data sampai tahapan terakhir yaitu eksperimen maka dapat disimpulkan bahwa algoritma C4.5 (*Decision Tree*) memiliki akurasi lebih tinggi dari k-NN menggunakan semua atribut dengan akurasi 98% sedangkan k-NN memiliki akurasi 83%. Sementara untuk penggunaan menggunakan fitur berdasar heatmap, algoritma C45 memiliki akurasi 88% lebih tinggi dari pada algoritma k-NN dengan akurasi 84%.Pada algoritma k-NN dengan mencari jumlah k yang optimal adalah 4 dengan akurasi 88%.

5. KESIMPULAN

Berdasarkan 16 atribut yang terdapat dalam dataset diabetes yaitu umur, *Alopecia*, *Gender*, *Polyuria*, Polydipsia, *sudden weight loss*, *weakness*, *Polyphagia*, *Genital thrush*, *Irritability*, *delayed healing*, *partial paresis*, *Itching*, *visual blurring*, *muscle stiffness*, dan Obesitas dapat dijadikan sebagai data untuk klasifikasi penderita penyakit diabates. Penelitian ini menggunakan dua algoritma yaitu Algoritma C4.5 dan k-NN untuk pengklasifikasian seseorang terkena penyakit diabetes atau tidak.

Total data sebanyak 520 dibagi menjadi 80% atau 416 data sebagai data *training* dan 20% atau 104 sebagai data *testing*. Dari hasil Pengujian diperoleh algoritma C45 (*Decision Tree*) memiliki akurasi lebih tinggi dari k-NN menggunakan semua atribut dengan akurasi 98% sedangkan k-NN memiliki akurasi 83%. Sementara untuk penggunaan menggunakan fitur berdasar heatmap, algoritma C45 memiliki akurasi 88% lebih tinggi dari pada algoritma k-NN dengan akurasi 84%.Pada algoritma k-NN dengan mencari jumlah k yang optimal adalah 4 dengan akurasi 88%.

DAFTAR PUSTAKA

[1] Kemenkes, "Pusat Data Dan Informasi Kementerian Kesehatan RI." 2021.
[2] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4 . 5," J. Sist. Komput. dan Kecerdasan Buatan, vol. Volume IV, 2020.
[3] M. A. A. K. Indrayanti, Devi Sugianti, "Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus," Pros. SNATIF Ke -4, pp. 823–829, 2017.
[4] D. T. Larose and C. D. Larose, Data Mining And Predictive Analytics. John Wiley & Sons, Inc., Hoboken, New Jersey, 2015.
[5] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: A review," Procedia Manuf., vol. 35, pp. 698–703, 2019, doi: 10.1016/j.promfg.2019.06.011.
[6] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition. 2020.
[7] D. T. Larose, Data Mining Methods And Models, vol. 28. John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.
[8] 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.
[9] F. D. Astuti and M. Guntara, "Analisis Performa Algoritma K-NN Dan C4.5 Pada Klasifikasi Data Penduduk Miskin," vol. 2, no. 2, 2018.
[10] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," Indones. J. Data Sci., vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.
[11] S. Iskandar, N. R. Refisis, and B. A. Ginting, "Metode Naive Bayes Classifer Dalam Penentuan Penerima Beasiswa Bidikmisi Di Universitas Negeri Medan," Karismatika, vol. 7, no. 1, pp. 10–23, 2021.
[12] S. Anggraini, S. Defit, and G. W. Nurcahyo, "Analisis Data Mining Penjualan Ban Menggunakan Algoritma C4. 5," *J. Ilmu Tek. Elektro …*, vol. 5, pp. 0–7, 2018.
[13] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on C4.5 algorithm for online voltage stability assessment," Int. J. Electr. Power Energy Syst., vol. 118, no. December 2019, p. 105793, 2020, doi: 10.1016/j.ijepes.2019.105793.
[14] I. W. Saputro and B. W. Sari, "Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa," Creat. *Inf. Technol. J.*, vol. 6, no. 1, p. 1, 2019, doi: 10.24076/citec.2019v6i1.178.