Klasifikasi Diabetes Menggunakan Metode *K-Nearest Neighbor*

(*Diabetes Classification using K-Nearest Neighbors*)

Ivan Andrianto[1], Ida Ayu Vigi Meidhyana Putri [2], Ni Nyoman Citariani Sumartha[3]
[1]Dept Informatics Engineering, Mataram University
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: ivandrian2024@gmail.com, idaayuvigi03@gmail.com, citariani359o@gmail.com

Diabetes mellitus is a metabolic disorder caused by the pancreas not producing enough insulin or the body unable to use the insulin it produces effectively. This disease is a disease that is suffered by many people and is classified as a deadly disease because it can cause several risks such as a stroke, heart disease, blindness, kidney failure and death. To predict diagnostically whether someone has diabetes or not, we classify it using the K-Nearest Neighbor (K-NN). K-NN is a method that looks for groups of objects in training data that are most similar to objects in training data that are most similar to objects in new data or testing data. This study used 9 attributes namely Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI (Body Mass Index), Diabetes Pedigree Function, Age and Outcome with 768 data from PIMA Indians Database to calculate accuracy, precision, recall and f-Score with a value of K = 1 to K = 100. The results of the calculation of the performance obtained the best K value is 34 where the accuracy value is 79.2%, precision 87%, recall 49%, and f-Score is 62.7%.

Keywords: Diabetes mellitus, K-NN, PIMA Indians Database, Training Data, Testing Data

I. PENDAHULUAN

Diabetes Mellitus atau dapat disebut dengan diabetes merupakan penyakit gangguan metabolik menahun akibat pankreas tidak memproduksi cukup insulin atau tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif. Diabetes mellitus akan meningkatkan gula darah dalam tubuh sehingga terjadi penyakit komplikasi yang dapat menyebabkan beberapa resiko seperti stroke, penyakit jantung, kebutaan, gagal ginjal, dan kematian [1].

Menurut Ahmed et al (2012), penyakit diabetes tergolong penyakit yang mematikan dan meningkatkan gula darah [2]. Penyakit ini dibagi menjadi 2 jenis yaitu diabetes tipe 1 dan tipe 2. Diabetes mellitus tipe 1 terjadi karena kerusakan dari tubuh menghasilkan insulin dan tubuh tidak dapat menggunakan insulin sebagaimana mestinya. Diabetes mellitus tipe 2 menghasilkan kelas dari resistensi insulin di mana sel-sel tidak bisa menggunakan insulin dalam proporsi yang tepat [3].

Pada tahun 2013, jumlah pasien diabetes mencapai 382 juta jiwa dan diperkirakan akan mencapai 595 juta jiwa pada tahun 2035. Sejak tahun 1965, orang Indian Pima yang tinggal di Komunitas Indian Sungai Gila di Arizona Selatan, AS telah berpartisipasi dalam studi longitudinal tentang diabetes dan komplikasinya. Suku ini memiliki prevalensi diabetes tertinggi yang dilaporkan di dunia (50% pada usia 35 tahun). Oleh karena itu, peneliti mencoba mengklasifikasikan seseorang terkena diabetes dan seseorang non-diabetes berdasarkan faktor (variabel) prediktor yang menurut World Health Organization sebagai kriteria penderita diabetes [4].

Sebagai contoh pada bidang medis, aplikasi klasifikasi dapat digunakan untuk klasifikasi tingkat penyakit yang diderita oleh seorang pasien sehingga memudahkan dokter dalam memberikan solusi terapi yang tepat. Untuk memecahkan masalah klasifikasi, berbagai macam metode dalam *Machine Learning* dapat diterapkan. Ketepatan dalam pengklasifikasian objek sangat penting, metode klasifikasi yang baik adalah metode yang menghasilkan kesalahan yang kecil, diantara metode klasifikasi yang telah ada yaitu metode *K-Nearest Neighbor* (K-NN) [5].

Salah satu penelitian yang berjudul "Klasifikasi Penyakit Jantung Menggunakan Metode *K-Nearest Neighbor*" melakukan pengklasifikasian dengan menggunakan 1000 data pasien di mana 90% dari data tersebut digunakan sebagai data *training* dan 10% sisanya digunakan sebagai data *testing*. Hasil pengklasifikasian menggunakan metode ini dengan simulasi K = 3 hingga K = 9 didapatkan bahwa nilai K = 6 memiliki nilai akurasi paling baik yaitu sebesar 85% dengan nilai presisinya sebesar 78%, *recall* sebesar 93% dan *F-Score* sebesar 85% [6].

Berdasarkan penelitian sebelumnya yang menggunakan metode *K-Nearest Neighbor* (KNN) didapatkan hasil akurasi yang baik. Oleh karena itu, kami melakukan penelitian dengan judul "Klasifikasi Diabetes Menggunakan Metode *K-Nearest Neighbor*" untuk mengidentifikasikan penyakit diabetes yang diharapkan dapat membantu tenaga medis dalam mengidentifikasi penyakit diabetes dengan menggunakan dataset Pima Indians Diabetes Database.

II. TINJAUAN PUSTAKA

A. Penelitian Terkait

Beberapa penelitian terkait yang pernah dilakukan oleh Ikhsan dkk dengan judul "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode *K-Nearest Neighbor*" berhasil melakukan pengklasifikasian

menggunakan dataset kanker payudara dengan 30 atribut dan 569 data yang menghasilkan tingkat akurasi sebesar 93% dengan menggunakan nilai K = 3 [7].

Penelitian yang pernah dilakukan oleh Olha Musa dan Alang pada tahun 2017 dengan judul "Analisis Penyakit Paru-paru Menggunakan Algoritma *K-Nearest Neighbors* Pada Rumah Sakit Aloei Saboe Kota Gorontalo" berhasil melakukan pengklasifikasian dan mampu mendeteksi penyakit paru-paru dengan tingkat akurasi sebesar 91.90% [8].

Selanjutnya penelitian yang pernah dilakukan oleh Hasran dengan judul "Klasifikasi Penyakit Jantung Menggunakan Metode *K-Nearest Neighbor*" berhasil melakukan pengklasifikasian dengan tingkat akurasi sebesar 85% yang menggunakan nilai K = 6. Data yang digunakan pada penelitian ini yaitu 1000 data pasien di mana 90% dari data tersebut digunakan sebagai data *training* dan 10% sisanya digunakan sebagai data *testing* [6].

B. *Diabetes Mellitus*

Diabetes Mellitus merupakan penyakit yang berisiko kematian tinggi. Penyakit *diabetes mellitus* terjadi ketika produksi insulin dalam tubuh tidak memadai atau tubuh tidak dapat memproduksi insulin dengan tepat. Pada umumnya penderita *diabetes mellitus* disebut terjadi ketika kadar gula darah berada di atas normal [9]. Penyakit *diabetes mellitus* sering mengancam kesehatan semua orang di seluruh dunia. Penyakit *diabetes mellitus* disebut juga penyakit komplikasi karena memiliki banyak gejala yang dapat mengakibatkan penyakit lainnya ataupun kematian [10].

B.1. Jenis penyakit *diabetes mellitus*

Penyakit *diabetes mellitus* dibagi menjadi 2 jenis yaitu:

- *Diabetes mellitus* tipe 1

  Penyakit *diabetes mellitus* tipe 1 biasanya disebut insulin *dependent*. *Diabetes mellitus* tipe 1 ini terjadi pada usia muda di bawah 30 tahun. Seseorang yang menderita *diabetes mellitus* tipe 1 perlu dilakukan suntik insulin. Suntik insulin dilakukan karena glukosa darah dalam tubuh tidak dapat memproduksi insulin sebagaimana mestinya [11].
- *Diabetes mellitus* tipe 2

  Penyakit *diabetes mellitus* tipe 2 biasanya disebut *non*-insulin *dependent* yang ditandai dengan resistensi insulin dan gangguan sekresi insulin. Tipe ini sering diderita oleh seseorang yang berusia di atas 40 tahun. Hal ini terjadi ketika tubuh manusia tidak dapat secara aktif menggunakan insulin yang dihasilkan oleh tubuh. Biasanya disebabkan faktor keturunan, obesitas, kurang aktivitas, penyakit lain dan usia [1].

C. Pima Indians Diabetes Database

Pima Indians Diabetes Database merupakan dataset diabetes yang umum digunakan para peneliti, karena data ini sudah tervalidasi dengan baik dapat memprediksi timbulnya penyakit diabetes. Data tersedia secara bebas pada UCI *Machine Repository Standard Dataset* untuk tujuan penelitian yang mencakup data penderita diabetes, dan *non*-diabetes, serta variabel-variabel indikator penyakit diabetes. Dataset ini berisi 768 contoh yang berbeda dan semua pasien dalam dataset adalah perempuan dengan kisaran umur 21 tahun dari keturunan India Pima [4].

Terdapat 8 variabel prediktor yang dipilih untuk memprediksikan timbulnya penyakit diabetes pada perempuan di Pima India. Variabel tersebut dipilih karena sudah dilakukan penelitian bahwa secara signifikan merupakan faktor resiko penyakit diabetes. Berikut penjelasan mengenai setiap variabel prediktornya dan hubungannya terhadap penyakit diabetes [4]:

- Variabel prediktor:
  1. *Pregnancies:* Banyaknya yang melahirkan.
  2. *Glucose*: Konsentrasi plasma glukosa.
  3. *Blood Pressure*: Tekanan darah diastole (mmHg).
  4. *Skin Thickness*: Ketebalan lipatan kulit trisep (mm).
  5. *Insulin*: Tergantung pada insulin yang dimiliki pasien.
  6. BMI (*Body Mass Index*): Index massa tubuh (Berat badan dalam kg/Tinggi badan dalam m^2).
  7. *Age*: Umur (tahun).
  8. *Diabetes Pedigree Function*: Fungsi yang menilai kemungkinan diabetes berdasarkan riwayat keluarga.
- Variabel hasil:
  1. *Outcome*: Kode kelas variabel yang terdiri dari dua kategori sebagai target (0 *or* 1), 0 untuk yang tidak mengidap diabetes, dan 1 untuk yang mengidap diabetes.

D. *Data Mining*

Data Mining adalah serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data. *Data mining* akan memecahkan masalah dengan menganalisis data yang telah ada dalam basis data. *Data mining*, sering juga disebut *knowledge discovery in database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan pola keteraturan, pola hubungan dalam set data berukuran besar. Hasil keluaran dari *data mining* ini dapat dijadikan untuk memperbaiki pengambilan keputusan di masa depan. Dalam penelitian ini kami akan memanfaatkan *data mining* untuk mengklasifikasikan data penyakit diabetes di mana hasil

keluarannya akan dimanfaatkan untuk keperluan pendeteksian seseorang yang mengidap penyakit diabetes [13].

E. Klasifikasi

Klasifikasi merupakan pengelompokkan sebuah objek ke dalam kelas tertentu. Berbagai kasus yang berkaitan dengan pengelompokkan objek dapat diselesaikan dengan menerapkan teknik-teknik klasifikasi [5].

Algoritma klasifikasi menggunakan data *training* untuk membuat sebuah model. Model yang sudah dibangun tersebut kemudian digunakan untuk memprediksi label kelas data baru yang belum diketahui. Dalam proses klasifikasi terdapat banyak algoritma yang telah dikembangkan oleh para peneliti seperti *K-Nearest Neighbor, Artifical Neural Network, Support Vector Machine, Decision Tree, Naïve Bayes Classifier* dan lain sebagainya. Prinsip masing-masing algoritma tersebut sama, yaitu melakukan suatu pelatihan sehingga di akhir pelatihan model dapat memprediksi setiap vektor masukan ke label kelas *output* dengan tepat [5].

Salah satu pengukur kinerja klasifikasi adalah tingkat akurasi. Sebuah sistem dalam melakukan klasifikasi diharapkan dapat mengklasifikasi semua set data dengan benar, tetapi tidak dipungkiri bahwa kinerja suatu sistem tidak bisa 100% akurat.

F. *K-Nearest Neighbor* (KNN)

Algoritma K-NN adalah algoritma yang menentukan nilai pada jarak pengujian data *testing* dengan data *training* berdasarkan nilai terkecil dari nilai ketetanggaan terdekat. Tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan *training samples* [5].

K-Nearest Neighbor (K-NN) merupakan metode yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data lain. Nilai K pada *K-Nearest Neighbor* (K-NN) berarti K-data terdekat dari data uji [5].

Adapun tahapan dari algoritma *K-Nearest Neighbor* (K-NN) dijelaskan sebagai berikut:

a. Menyiapkan data *training* dan data *testing*.
b. Menentukan nilai K.
c. Menghitung jarak data *testing* ke setiap data pelatihan.

   Data pelatihan dihitung menggunakan rumus perhitungan jarak *Euclidean* pada Persamaan (1).

$$d(x1, x2) = \sqrt{\sum_{i=1}^{n}(x1i - x2i)^2} \quad (1)$$

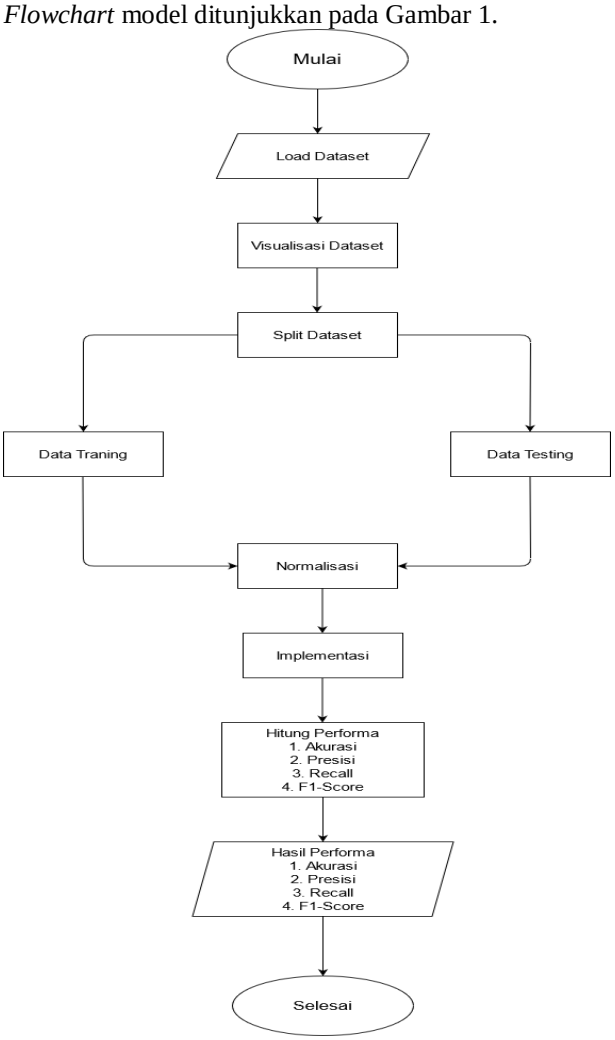
   Keterangan:
   x1i = Data *Training*
   x2i = Data *Testing*
d. Menetukan nilai K data *training* yang memiliki jarak terdekat dengan data *testing*.
e. Memeriksa label dari K data *training*.
f. Menentukan label yang frekuensinya paling banyak.
g. Memasukkan data *testing* ke dalam kelas dengan frekuensi paling banyak.
h. Kondisi berhenti.

K-Nearest Neighbor (KNN) memiliki kelebihan yaitu dapat menghasilkan data yang kuat atau jelas dan efektif jika digunakan pada data besar. *K-Nearest Neighbor* juga memiliki kekurangan yaitu membutuhkan nilai K, jarak dari data percobaan tidak jelas dengan tipe jarak yang digunakan, untuk memperoleh hasil yang terbaik maka harus menggunakan semua atribut atau hanya satu atribut yang telah pasti. Pemilihan nilai K (jumlah data/tetangga terdekat) ditentukan oleh peneliti. Pemilihan nilai K akan mempengaruhi tingkat akurasi prediksi yang dikerjakan.

III. METODE PENELITIAN

A. Rancangan / Model

Rancangan atau model pada penelitian ini seiring dengan alur proses metode K-NN di mana dimulai dari mengumpulkan dataset, visualisasi dataset, *split* atau membagi dataset, implementasi metode K-NN hingga menghitung performa metode K-NN pada dataset tersebut. *Flowchart* model ditunjukkan pada Gambar 1.



Gambar 1. *Flowchart* model penelitian metode KNN

B. Sampel dan Data

Data yang diolah pada penelitian ini menggunakan dataset dari Pima Indians Diabetes yang terdiri dari beberapa variabel prediktor medis (independen) dan satu variabel target (dependen). Variabel independen meliputi *Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin,* BMI, *Diabetes Pedigree Function, Age*. Variabel dependen berupa hasil (*Outcome*) yang terdiri dari 2 kategori sebagai target (0 *or* 1), 0 untuk yang tidak mengidap diabetes, dan 1 untuk yang mengidap diabetes.

Adapun jumlah dataset yang kami gunakan yaitu 768 data dengan 80% (614 data) digunakan sebagai data *training* dan 20% (154 data) digunakan sebagai data *testing*.

C. Evaluasi dan Hasil Klasifikasi

Pada penelitian ini, dilakukan evaluasi terhadap hasil klasifikasi dengan menghitung nilai *True Positive, True Negative, False Positive,* dan *False Negative*. *True Positive* adalah banyaknya hasil klasifikasi benar untuk suatu kelas yang bernilai *positive. True Negative* adalah banyaknya hasil klasifikasi benar untuk suatu kelas yang bernilai *negative*. *False Positive* adalah banyaknya hasil klasifikasi salah untuk suatu kelas yang bernilai *positive*. *False Negative* adalah banyaknya hasil klasifikasi salah untuk suatu kelas yang bernilai *negative.* Keempat nilai tersebut dapat dihitung dengan *confusion matrix* pada Tabel I [14]. Nilai-nilai inilah yang kemudian digunakan ntuk menghitung parameter-parameter evaluasi hasil klasifikasi.

TABEL I. CONFUSION MATRIX

Hasil Klasifikasi / **Kelas sebenarnya**	**Positif**	**Negatif**
Positif	TP	FN
Negatif	FP	TN

Akurasi dapat diartikan sebagai proporsi dari dua kelas (positif dan negatif) dari jumlah total kelas yang diujikan. Berikut Persamaan (2) adalah rumus yang digunakan untuk menghitung nilai akurasi.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Recall adalah proporsi dari kelas positif yang diklasifikasi dengan benar. Berikut Persamaan (3) adalah rumus yang digunakan untuk perhitungan nilai *recall.*

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

Presisi adalah proporsi dari kelas positif yang diklasifikasikan benar positif dibandingkan dengan keseluruhan hasil yang diklasifikasikan positif. Berikut Persamaan (4) adalah rumus yang digunakan untuk perhitungan nilai presisi.

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (4)$$

D. Normalisasi *Min-Max*

Pada penelitian ini, dilakukan normalisasi *min-max* untuk mengubah ukuran data dari rentang asli, sehingga semua nilai berada dalam kisaran 0 dan 1. Rumus perhitungan normalisasi *min-max* dapat dilihat pada Persamaan (5) [15].

$$V_{\text{norm}} = \frac{(\text{Vi}-\text{Vmin})}{(Vmax-Vmin)} \quad (5)$$

E. Langkah Pengujian

Tahap awal yang dilakukan adalah mengumpulkan data, data yang diperoleh adalah sebanyak 768 data, data yang kami gunakan dalam penelitian ini dapat di unduh pada:ftp.ics.uci.edu/pub/machine-learning-databases/pima-indians-diabetes/.

Tahap kedua adalah melakukan *split* data di mana 80% digunakan sebagai data *training* (data latih), dan 20% digunakan sebagai data *testing* (data uji). Pada tahap ini kami akan melatih pengklasifikasian kami menggunakan data latih kemudian menggunakan data uji sebagai *input* untuk menguji seberapa akurat pengklasifikasian kami dalam menebak hasil.

Tahap ketiga adalah melakukan normalisasi dengan tujuan memperoleh data yang memiliki ukuran yang lebih kecil untuk mewakili data asli dengan tidak menghilangan karakteristik dirinya karena kami tidak ingin terdapat sebuah fitur yang lebih dominan dibanding fitur lainnya.

Tahap keempat adalah menghitung jarak antar data dengan menggunakan rumus *Euclidean Distance* seperti pada Persamaan (1).

Tahap kelima adalah menentukan nilai K yang akan digunakan. K merupakan jumlah tetangga yang ingin kami gunakan sebagai acuan klasifikasi. Nilai yang yang akan digunakan untuk klasifikasi tidak boleh sangat kecil maupun sangat besar. Jika menggunakan nilai K yang sangat kecil maka pengklasifikasi akan mengalami *overfitting* karena tidak mempertimbangkan cukup tetangga. Dan jika menggunakan nilai K yang sangat besar, pengklasifikasi akan mengalami *underfitting* karena tidak cukup memperhatikan detail saat melakukan pelatihan. Nilai K yang disarankan adalah bilangan ganjil untuk menghindari ambiguitas. Misalkan nilai K adalah bilangan genap dan didapatkan hasil frekuensi yang sama antar label *output* maka pengklasifikasi akan bingung dalam mengklasifikasikan hasil.

Tahap terakhir adalah mengklasifikasi data uji berdasarkan tetangganya. Setelah didapatkan nilai jaraknya maka kami akan memilih data dengan jarak terpendek sebanyak nilai K yang diinginkan. Kemudian dari tetangga yang dipilih, lihat label data yang paling dominan dari tetangga tersebutlah yang menjadi hasil prediksi dari data uji. Sehingga setelah itu kami dapat membandingkan hasil prediksi dengan hasil sebenarnya dari label data uji sehingga dapat dilakukan evaluasi terhadap kinerja pengklasifikasian kami.

IV. HASIL DAN PEMBAHASAN

Sebelum melakukan proses perhitungan K-NN, disiapkan dataset diabetes berjumlah 768 data dengan 9 atribut yang digunakan untuk melakukan klasifikasi.

TABEL II. TAMPILAN 5 DATA TERATAS

	0	1	2	3	4
Pregnancies	6	1	8	1	0
Glucoses	148	85	183	89	137
Blood Pressure	72	66	64	66	40
Skin Thickness	35	29	0	23	35
Insulin	0	0	0	94	168
BMI	33.6	26.6	23.3	28.1	43.1
Diabetes Pedigree Function	0.627	0.351	0.672	0.167	2.288
Age	50	31	32	21	33
Outcome	1	0	1	0	1

Tabel II merupakan tampilan 5 data teratas dari dataset diabetes yang kami gunakan pada penelitian ini. Kemudian setelah itu kami mengelompokkan data yang berjumlah 768 data ke dalam 2 kategori, yaitu data *training* dengan jumlah 614 data dan data *testing* dengan jumlah 154 data. Data *training* merupakan kelompok data yang akan digunakan pengklasifikasi dalam mempelajari dan memahami informasi pada data tersebut sehingga pengklasifikasi nantinya dapat memprediksi hasil *output* berdasarkan pembelajaran pola informasi yang telah dilakukan. Data *testing* merupakan kelompok data yang akan diujicobakan untuk diprediksi hasil *output*-nya dan diharapkan agar hasil prediksi dan label sebenarnya sesuai sehingga dapat disimpulkan bahwa kinerja pengklasifikasi sangat baik.

Selanjutnya, kami menghitung jarak antar data dengan menggunakan rumus *Euclidean Distance* seperti pada Persamaan (1). Lalu setelah itu dilakukan perhitungan performa untuk mendapatkan nilai akurasi, *recall*, presisi, dan F1-*Score* dengan menggunakan nilai K = 3.

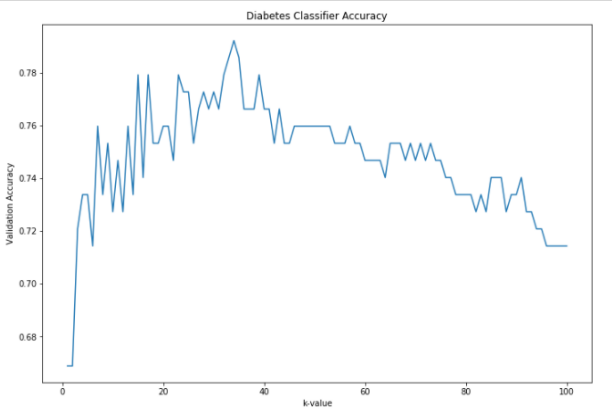
Hasil akurasi yang kami dapatkan pada penelitian ini dengan menggunakan K = 3 adalah 0.712. Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual [3].

Hasil *recall* yang kami dapatkan pada penelitian ini dengan menggunakan K = 3 adalah 0.613. *Recall* didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia [3].

Hasil presisi yang kami dapatkan pada penelitian ini dengan menggunakan K= 3 adalah 0.607. Presisi didefinisikan sebagai rasio dari item relevan yang dipilih terhadap semua item yang terpilih [3].

Hasil F1-*Score* yang kami dapatkan pada penelitian ini dengan menggunakan K = 3 adalah 0.613. F1-*Score* adalah *harmonic mean* antara nilai presisi dan *recall*.

Kemudian, setelah mendapatkan hasil nilai akurasi, *recall*, presisi, dan F1-*Score* dengan menggunakan nilai K = 3, kami mencoba untuk membandingkan hasil nilai akurasi terbaik yang terjadi ketika rentang nilai K = 1 hinnga K = 100. Berikut merupakan grafik hasil nilai akurasi yang kami dapatkan:



Gambar 2. Grafik hasil akurasi K = 1 hingga K = 100

Dari Gambar 2 dapat dibuktikan bahwa perbedaan nilai K akan memengaruhi hasil akurasi yang didapatkan. Hasil akurasi terbaik dari grafik terlihat ketika nilai K dalam rentang 30 hingga 40. Namun, tidak dapat diketahui secara spesifik ketika nilai K berapakah itu. Untuk itu, kami membuat tabel hasil akurasi, presisi, *recall*, dan F1-*Score* dalam bentuk *DataFrame* sehingga dapat dilihat spesifik hasil evaluasi data terbaik dengan mudah.

	k	accuracy	precision	recall	f1-score
0	1	0.668831	0.534483	0.563636	0.548673
1	2	0.668831	0.562500	0.327273	0.413793
2	3	0.720779	0.607143	0.618182	0.612613
3	4	0.733766	0.718750	0.418182	0.528736
4	5	0.733766	0.659091	0.527273	0.585859

Gambar 3. 5 data teratas berdasarkan Gambar 2

Gambar 3 merupakan tampilan tabel hasil perhitungan nilai akurasi, presisi, *recall*, dan f1-*score* yang menampilkan 5 data teratas dari grafik pada Gambar 2.

	k	accuracy	precision	recall	f1-score
2	3	0.720779	0.607143	0.618182	0.612613
14	15	0.779221	0.769231	0.545455	0.638298
33	34	0.792208	0.870968	0.490909	0.627907

Gambar 4. Data terbaik dengan nilai K = 3, 15, 34

Dari Gambar 4 dapat terlihat hasil nilai akurasi dan presisi terbaik didapatkan ketika nilai K = 34 dengan nilai akurasi sebesar 0.792208 atau 79.2% dan nilai presisi sebesar 0.870968 atau 87%. Lalu nilai *recall* terbaik didapatkan ketika nilai K = 3 sebesar 0.618182 atau 61.8%, dan nilai f1-*score* terbaik didapatkan ketika nilai K = 15 sebesar 0.638298 atau 63.8%.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, metode K-NN mampu mengklasifikasikan tingkat akurasi seseorang yang mengidap diabetes menggunakan 768 data. Untuk mendapatkan tingkat akurasi terbaik, kami menggunakan nilai K = 1 hingga K = 100 yang kemudian didapatkan hasil akurasi terbaik sebesar 79.2% pada K = 34. Selain nilai akurasi terbaik, didapatkan juga hasil perhitungan presisi terbaik ketika nilai K = 34 dengan nilai presisi sebesar 87%. Lalu nilai *recall* terbaik ddapatkan ketika nilai K = 3 sebesar 61.8%, dan nilai F1-*Score* terbaik didapatkan ketika nilai K = 15 sebesar 63.8%.

B. Saran

Penelitian ini mungkin masih terdapat beberapa kekurangan sehingga masih dapat dikembangkan lagi dengan menghilangkan *outlier* atau pencilan data dari tiap atribut atau dapat menggunakan metode lainnya untuk menghasilkan nilai akurasi yang lebih baik.

UCAPAN TERIMA KASIH

Terima kasih kami ucapkan kepada Bapak Gibran Satya Nugraha, S.Kom., M.Eng. selaku dosen pengampu mata kuliah Bioinformatika dan Bapak Ramaditia Dwiyansaputra, S.T., M.Eng. selaku dosen pengampu mata kuliah Kecerdasan Buatan yang telah memberikan pengetahuan mengenai algoritma KNN dan telah membimbing kami mendapatkan pengalaman dalam menulis sebuah jurnal dengan baik hingga akhirnya kami dapat menyelesaikan jurnal penelitian ini.

DAFTAR PUSTAKA

[1] V. VijayanV and A. Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus," *Int. J. Comput. Appl.*, vol. 95, no. 17, pp. 12–16, 2014, doi: 10.5120/16685-6801.

[2] K. Ahmed, T. Jesmin, and U. Fatima, "Intelligent and Effective Diabetes Risk Prediction System Using Data Mining," *Orient. J. Comput. Sci. Technol.*, vol. 5, no. 2, pp. 215–221, 2012.

[3] S. Peter, "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus," *Bonfring Int. J. Data Min.*, vol. 4, no. 2, pp. 7–11, 2014.

[4] I. L. MULYAHATI, "KNN-Pima Indians Diabetes Database dengan Menggunakan Python," *medium.com*, 2019. https://medium.com/@indiraluthfianam/knn-pima-indians-diabetes-database-dengan-menggunakan-python-part-1-2-8812372c0460 (accessed Dec. 01, 2020).

[5] F. Yunita, "Diabetes Mellitus Menggunakan Metode K-Nearest Neighbor (K-Nn)," *Bappeda*, vol. 2, pp. 223–230, 2016.

[6] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020.

[7] I. N. Atthalla, A. Jovandy, and H. Habibie, "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K-Nearest Neighbor," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 978–979, 2018.

[8] O. Musa and Alang, "Analisis Penyakit Paru-Paru Menggunakan Algoritma," *Ilk. J. Ilm.*, vol. 9, pp. 348–352, 2017.

[9] A. Iyer, J. S, and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 1, pp. 01–14, 2015, doi: 10.5121/ijdkp.2015.5101.

[10] K. M. Juliyet, L. C., & Amanullah, "The Surveillance on Diabetes Diagnosis Using Data Mining Techniques A case study in the Medical Diagnosis," *IJSART*, vol. 1, no. 4, 2015.

[11] S. Sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi, "Comparison of Data Mining Algorithms in the Diagnosis of Type Ii Diabetes," *Int. J. Comput. Sci. Appl.*, vol. 5, no. 5, pp. 1–12, 2015, doi: 10.5121/ijcsa.2015.5501.

[12] Y. Aisyah, F. Bimantoro, and B. Irmawati, "Sistem Pakar Diagnosa Penyakit Gigi Dengan Metode Bayesian Network Berbasis Website," *J-Cosine*, vol. 3, no. 2, pp. 137–143, 2019.

[13] A. Rohman, "MODELALGORITMA K-NEARESTNEIGHBOR(K-NN)UNTUK PREDIKSI KELULUSAN MAHASISWA," *NEO Tek.*, vol. 1, no. 1, 2015, doi: http://dx.doi.org/10.37760/neoteknika.v1i1.350.

[14] and I. G. . P. S. W. F. Bimantoro, Nurhalimah, "GLCM DAN MOMENT INVARIANT DENGAN TEKNIK PENGKLASIFIKASIAN LINEAR DISCRIMINANT ANALYSIS (LDA)," *J-Cosine*, vol. 2, no. 1, pp. 173–183, 2020.

[15] Y. H. Agus Ambarwari, Qadhli Jafar Adrian, "Jurnal Resti," *RESTI*, vol. 4, no. 1, pp. 117–122, 2017.