Journal of Information Technology and Computer Science (INTECOMS)
Volume 6 Nomor 1, Juni 2023
e-ISSN : 2614-1574
p-ISSN : 2621-3249

IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA NAIVES BAYES DAN K-NEAREST NEIGHBOR

IMPLEMENTATION OF DATA MINING TO PREDICT DIABETES DISEASE USING NAIVES BAYES AND K-NEAREST NEIGHBOR ALGORITHMS

Fitrokh Nur Ikhromr[1)], Ipin Sugiyarto[2)*], Umi Faddillah[3)], Bibit Sudarsono[4)]
[1)2)]Universitas Nusa Mandiri, [3)4)]Universitas Bina Sarana Informatika, Indonesia
[1)]fitrohnur260@gmail.com , [2)]ipin.isy@nusamandiri.ac.id , [3)]umi.umf@bsi.ac.id, [4)]bibit.bbs@bsi.ac.id

ABSTRACT

Early handling of diabetes risk and paying attention to the factors that cause a person to have the potential for diabetes is important to be able to minimize cases of diabetes sufferers. Monitoring Pre-diabetic patients characterized by increasing certain parameters in medical record data features are an important dynamic part of this research. Data mining techniques in diabetes prediction are used to determine a patient at risk of diabetes more quickly and accurately. This research uses the Knowledge Discovery in Database model process which consists of several stages such as, Data Selection, Preprocessing, Transformation, Data Mining and Evaluation. Model testing using Navies Bayes and K-Nearest Neighbor algorithms has mixed evaluation results with several input datasets used. Evaluation results using 2000 diabetic patient datasets K-Nearest Neighbor produces 99% accuracy while Naives Bayes produces 75% accuracy. The following test using 30 test datasets with the K-Nearest Neighbor algorithm resulted in 53% accuracy while Navies Bayes resulted in 66% accuracy. Several experiments using training data sets and testing data sets conducted to see the best results of the performance of each algorithm to predict diabetic patients prove that the evaluation model using the K-Nearest Neighbor algorithm gets the best results..

Keywords: *K-Nearest Neighbor; Naïve Bayes; Prediction; Accuracy; Diabetes*

ABSTRAK

Penanganan dini risiko diabetes dan memperhatikan faktor penyebab seseorang memiliki potensi diabetes menjadi hal penting untuk dapat meminimalisir kasus penederita diabetes. Pemantauan pasien Pra-diabetes yang ditandai dengan meningkatkan parameter tertentu dalam fitur data rekam medis yang merupakan bagian dinamika penting dalam penelitian ini. Teknik penambangan data pada prediksi penyakit diabetes digunakan untuk menentukan seorang pasien risiko diabetes dengan lebih cepat dan akurat. Penelitian ini menggunakan proses model *Knowledge Discovery in Database* yang terdiri dari beberapa tahap seperti, *Data Selection*, *Preprocessing*, *Transformation*, *Data Mining* dan *Evaluation*. Pengujian model menggunakan algoritma Navies bayes dan K-Nearest Neighbor memiliki hasil evaluasi beragam dengan beberapa input dataset yang digunakan. Hasil evaluasi menggunakan 2000 set data pasien diabetes K-Nearest Neighbor menghasilkan akurasi sebesar 99% sedangkan Naives Bayes mengahsilkan akurasi sebesar 75%. Pengujian berikut menggunakan 30 set data uji dengan algoritma K-Nearest Neighbor menghasilkan akurasi 53% sedangkan Navies Bayes menghasilkan akurasi 66%. Beberapa percobaan dengan menggunakan set data training dan set data testing yang dilakukan untuk melihat hasil terbaik dari performa masing-masing algoritma untuk memprediksi pasien diabetes membuktikan bahwa model evaluasi menggunakan algoritma K-Nearest Neighbor mendapatkan hasil terbaik.

Kata Kunci: K-Nearest Neighbor; Naïve Bayes; Prediction; Accuracy; Diabetes

PENDAHULUAN

Diabetes yaitu gangguan metabolisme yang diidentifikasi melalui hiperglikemia yang diakibatkan oleh ketidak mampuan pankreas untuk mensekresi insulin, gangguan kerja insulin. Kondisi hiperglikemik kronis dapat menimbulkan kerusakan berkepanjangan dan tidak berfungsinya beberapa organ seperti mata, ginjal, jantung, dan pembuluh darah (Joshua Neumiller, 2020). Menurut *International Diabetes Federation* (IDF), 433 juta manusia di seluruh dunia menderita diabetes pada tahun 2019. Total itu diproyeksikan meningkat 578 juta pada 2030 dan 700 juta lagi pada 2045. Indonesia sendiri termasuk kedalam salah satu dari 10 negara dengan total penderita diabetes

terbesar di dunia di tahun 2019 (Bingga, 2021). Dengan melihat banyaknya total kasus diabetes, maka memerlukan tindakan awal untuk tindakan dini penyakit diabetes dengan melakukan prediksi.

Prediksi untuk penderita diabetes dapat ditemukan dengan mengumpulkan sejumlah besar data tentang penderita diabetes, menyimpannya dalam database, dan kemudian mengolah hasilnya dalam pola tertentu Untuk deteksi dini diabetes (Patwari, 2021). Berbagai penelitian yang tealah dilakukan untuk melakukan prediksi diabetes diantaranya dengan algoritma ID3 yang telah dilakukan (Wibawa, 2018). Selain itu prediksi tersebut juga dilakukan dengan menggunakan algoritma K-*Nearest Neighbor* (KNN) (Argina A. , 2020).

Naïve Bayes yaitu metode klasifikasi yang menggunakan perhitungan peluang. Penetapan kelas dari suatu data pada data dilakukan dengan membadingkan nilai peluang suatu sample ada di kelas yang berbeda. algoritma klasifikasi *Naïve Bayes* yaitu teknik pembelajaran *Bayesian* yang diketahui sangat bermanfaat dalam berbagai aplikasi. *Naïve Bayes* adalah metode *supervised learning*. teknik ini diketahui mempunyai tingkat akurasi yang baik dengan perhitungan sederhana (N. Maulidah, 2021).

Algoritme K-Nearest Neigbor (KNN) termasuk dalam grup pembelajaran berbasis instans. Algoritma ini merupakan teknik *lazy learning*. K-Nearest Neighbor (KNN) dikerjakan dengan menentukan kelompok K objek pada data latih yang terdekat dengan tujuan pada data baru atau data pengujian. Dibutuhkan metode klasifikasi sebagai metode yang ahli menemukan informasi. Algoritma ini bertindak berdasarkan jarak terdekat dari data latih ke data pengujian untuk menetapkan K-Nearest Neighbor. sesudah mengumpulkan K- Nearest Neighbors, lalu diambil sebagian besar K- Nearest Neighbors untuk dibuat prediksi dari sampel uji (Argina A. , 2020).

Berdasarkan latar belakang permasalahan tersebut Algoritma K-Nearest Neighbor dan Naives Bayes sama-sama digunakan dalam penyelesaian model klasifikasi. Tetapi dalam hal ini memiliki beberapa karakteristik penyelesaian yang berbeda sehingga dibutuhkan pengujian lebih lanjut untuk mengetahui performa algoritma yang terbaik dalam menyelesaikan permasalah klasifikasi untuk kasus pencarian pola dalam dataset diabetes dengan jumlah data 2000 record, 9 atribut prediktor dan kelas dengan 2 outcome menjadi objek penelitian ini.

Diabetes dikenal sebagai kencing manis, adalah penyakit yang berhubungan dengan meningkatnya kadar gula darah dalam tubuh, terutama setelah makan. Salah satu ciri-ciri diabetes adalah meningkatnya kadar gula darah yang digambarkan di atas normal atau hipertensi (120 mg / dl atau 120 mg% atau lebih) (Rahayu P.T., 2022). Banyak faktor yang mempengaruhi orang menderita diabetes, beberapa diantaranya yaitu tekanan darah tinggi, kadar gula berlebih, berat badan, riwayat keturunan diabetes, usia, jumlah kehamilan seseorang, ketebalan lipatan kulit, dan jumlah kadar insulin dalam tubuh (Qatrunnada Refa Cahyani, 2022).

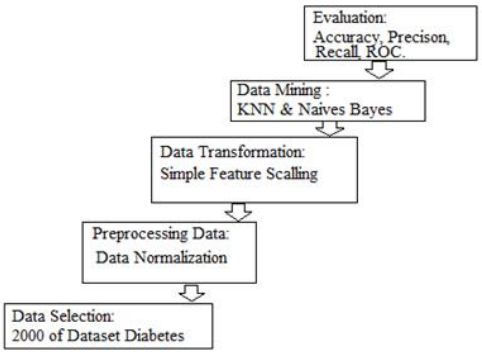
Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat. Data Mining didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, data mining dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering dan asosiasi. Proses dalam tahap data mining terdiri dari tiga langkah Utama, yaitu data Preparation Pada langkah ini, data dipilih, dibersihkan, dan dilakukan preprocessed mengikuti pedoman dan knowledge dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh (Argina A. M., 2020).

Dalam regresi logistik, peneliti ingin menguji prediktor atau sekumpulan prediktor dari ketergantungan dikotomis

variabel, seperti pasien hidup atau mati, atau pasien menanggapi atau tidak menanggapi pengobatan (Connely, 2020). Variabel independen dapat berupa nominal, ordinal (peringkat), interval, atau tingkat rasio (atau kontinu) data. Proses KDD secara garis besar seperti Data selection, Pre-processing/ Cleaning, Transformation, Data mining dan Interpretation/Evaluation (Novrizal Eka Saputra, 2016). Pengujian KNN dilakukan dengan menghitung jarak antara data testing dengan data training (Permadi. J., 2021). Metode lain untuk mengevaluasi hasil klasifikasi dari model yang berbeda-beda dapat menggunakan confusion matrix (F., 2018). Naïve Bayes adalah metode data mining yang sederhana dan mudah untuk diimplementasikan dibandingkan metode yang lain dalam konteks klasifikasi. Metode ini juga mampu mengolah data numeric dan teks (Hendrik, 2018). Confusion Matrix merupakan pengukuran performa buat permasalahan klasifikasi machine learning dimana keluaran bisa berbentuk 2 kelas ataupun lebih. Confusion Matrix merupakan tabel dengan 4 campuran berbeda dari nilai prediksi serta nilai aktual (Hozairi Hozairi, 2021).

METODE

Metode penelitian dilakukan dengan tahapan metode Knowledge Discovery in Data (KDD). Alur penelitian merupakan tahapan penelitian untuk menjelaskan langkah-langkah yang digunakan dalam melaksanakan penelitian. Langkah pertama ialah mengumpulkan datapasien diabetes dengan jumlah 2000 record, 9 atribut prediktor, dani kelas prediktor dengan 2 outcome di ambil dari sumber National Institute of Diabetes and Digestive and Kidney Diseases.



Gambar 1. KDD Methodology Modifikasi

Tahapan data *pre-processing* yaitu proses data mining yang utama dilakukan untuk memperoleh kualitas data sebelum melakukan proses klasifikasi seperti mentransformasikan data. Tahap transformasi data merupakan salah satu bagian penting dari tahap data *pre-processing* untuk mentransformasikan rentang nilai pada data tetapi tidak menghilangkan informasi dari data yang ada Hal ini diterapkan untuk mempermudah perhitungan nilai kolerasi setiap atribut terhadap kelas pada metode *Naïve Bayes* dan K-*Nearest Neighbor* pada tahap klasifikasi.

Pada penelitian ini untuk menganalisa performa model klasifikasi beberapa algoritma menggunakan aplikasi *orange*, untuk melakukan perbandingan dua algoritma data mining untuk menentukan algoritma terbaik dengan *accuracy* tinggi, dalam mengklasifikasikan data *Diabetes Prediction Using Logistic Regression*. Dalam penelitian ini menggunakan algoritma klasifikasi pada aplikasi *orange* berbentuk *Naïve Bayes* dan K-*Nearest Neighbor* yang telah penginputan data yang sudah proses sebelumnya. Lalu data akan diproses kedalam model klasifikasi.

Pada tahap melakukan proses pengujian algoritma klasifikasi, perlu dilakukan input data uji dalam melakukan proses klasifikasi untuk mengetahui hasil klasifikasi *Diabetes Prediction Using Logistic Regression*. Tahap berikutnya adalah membandingkan algoritma klasifikasi yang digunakan dengan

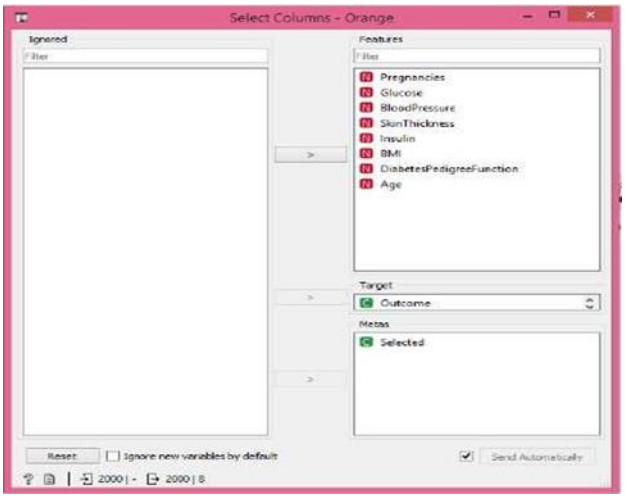
memakai *tes and score* yang digunakan untuk mengukur tingkat keberhasilan antara tiap tiap algoritma klasifikasi pada *orange*. kemudian akurasinya akan dievaluasi dengan menggunakan *confusion matrix* dan ROC *analysis.*

HASIL DAN PEMBAHASAN

a. Analisis Data

1) Data Selection Dataset

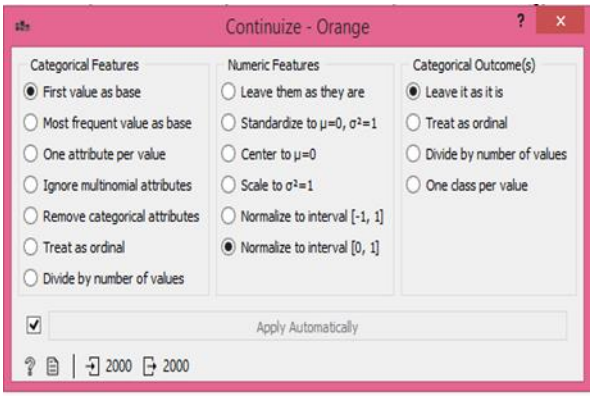
Pada proses data *selection* dataset *Diabetes Prediction Using Logistic Regression* ini tidak terdapat *missing value* di datanya jadi hanya dilakukan pemilihan data saja.



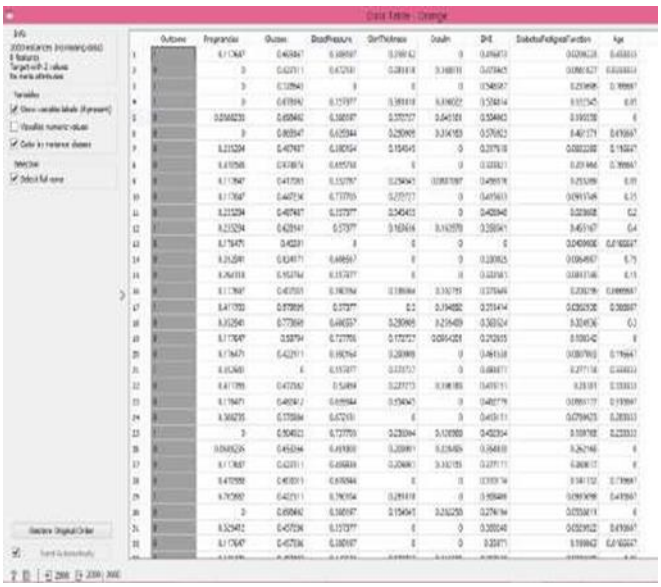
Gambar 2. Proses Pemilihan data

Proses pemilihan data pertama memilih kolom dimana *attribute* yang digunakan adalah *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, insulin, BMI, *DiabetesPedigreeFunction*,dan *Age* dengan *attribute* target adalah *Outcome*. Pada tahap *pre-processing* adalah tahap data mining yang dikerjakan untuk memperoleh kualitas data sebelum melakukan tahap klasifikasi. Pada proses ini di lakukan mengubahan data untuk mentransformasikan rentang nilai pada data tetapi tidak menghilangkan informasi dari data yang ada.

Gambar. 3 Proses Normalisasi Data

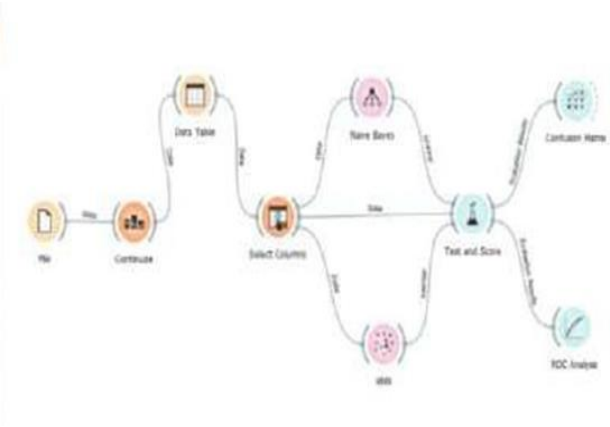


dapat dijelaskan dalam penelitian ini untuk menormalisasikan data menggunakan metode *simple feature scalling* dimana teknik normalisasi data yang akan menghasilkan rentang nilai dari 0 sampai 1.



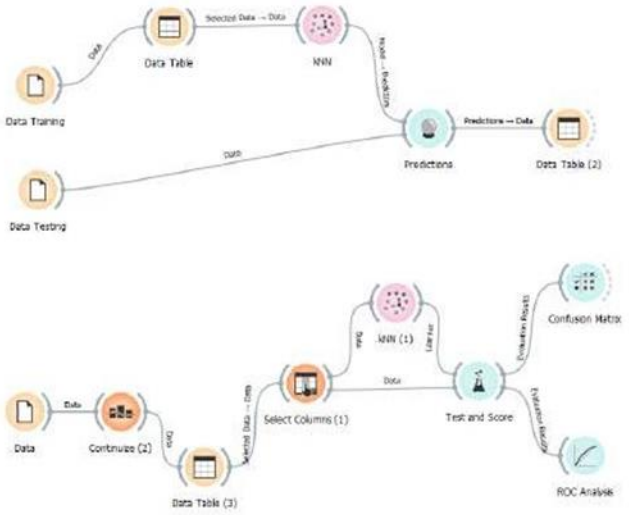
Gambar 4. Hasil Normalisasi Data Sample Feature Scalling

Hasil dari normalisasi *simple feature scalling* yang digunakan mengubah rentang nilai data dari 0 sampai 1, tetapi tidak menghilangkan informasi dari data yang ada. Pada proses data mining melakukan perbandingan dua algoritma data mining untuk menentukan algoritma yang terbaik melalui nilai *accuracy* yang terbaik, dalam mengklasifikasi data *Diabetes Prediction Using Logistic* bisa dilihat pada gambar berikut :



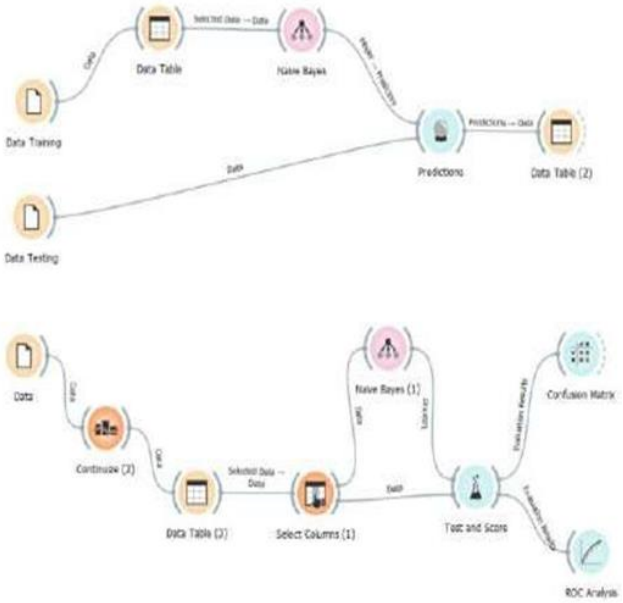
Gambar 5. Design Widget Model Klasifikasi Prediksi Diabetes

design *widget* menggunakan algoritma klasifikasi di aplikasi *orange* berupa *Naïve Bayes* dan K-*Nearest Neighbor* bahwa telah dimasukkan data yang sudah diproses sebelumnya. Lalu data diproses dalam algoritma klasifikasi.



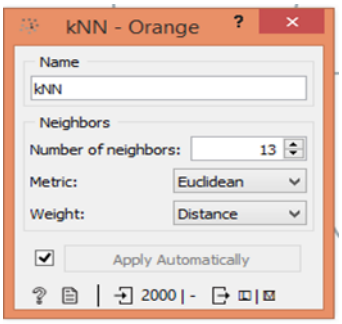
Gambar 6. Design Widget Model Klasifikasi Prediksi Diabetes K-Nearest Neighbor

perancangan widget yang menggunakan algoritma K-Nearest Neighbor dengan data yang yang tidak berlabel yang akan dijadikan data testing untuk pengujian klasifikasi prediksi diabetes. Algoritma K-Nearest Neighbor merupakan salah satu metode klasifikasi data mining, KNN mengklasifikasikan sekumpulan data berdasarkan data pembelajaran diberi label (B., 2020).



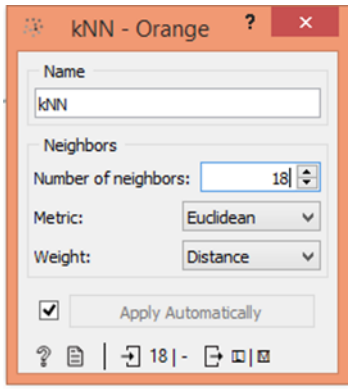
Gambar 7. Design Widget Model Klasifikasi Diabetes Naives Bayes

perancangan *widget* yang menggunakan algoritma *Naïve Bayes* dengan data yang yang tidak berlabel yang akan dijadikan data testing untuk pengujian klasifikasi prediksi diabetes. Naive Bayes merupakan metode yang membagi permasalahan ke dalam sebuah kelas-kelas berdasarkan ciri-ciri persamaan dan perbedaan dengan menggunakan statistik yang bisa memprediksi probabilitas sebuah kelas (Nurdiana et al., 2020).



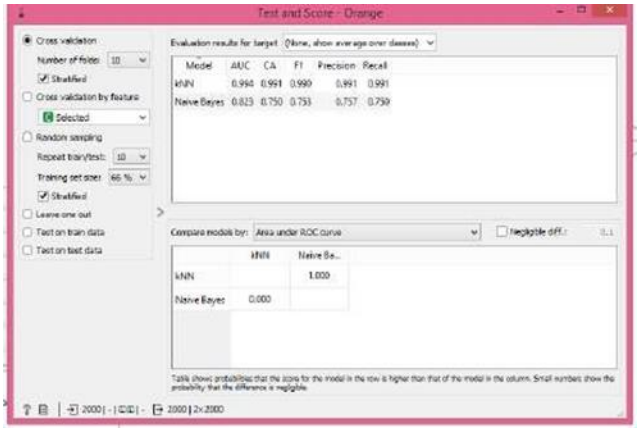
Gambar. 8 Penentuan nilai K untuk K-Nearest Neighbor

penentuan nilai K untuk metode klasifikasi K- *Nearest Neighbor* yang akan dipakai untuk mengklasifikasikan prediksi diabetes. dapat dijelaskan untuk nilai K untuk K-*Nearest Neighbor* yaitu K=13 dan memilih *matrix Euclidean* untuk mengetahui akurasi dengan cara menghitung *enclidean distance*, serta *weight* yang digunakan adalah *distance* karena akan menghitung jarak dengan yang paling terdekat



Gambar 9. Penentuan Nilai K untuk K-Nearest Neighbor 30 Data

penentuan nilai K dengan meggunakan 30 data untuk metode klasifikasi K-*Nearest Neighbor* yang akan dipakai untuk mengklasifikasikan prediksi diabetes. dapat dijelaskan untuk nilai K pada K- *Nearest Neighbor* adalah K=18 dan memilih *matrix Euclidean* untuk mengetahui akurasi dengan cara menghitung *enclidean distance*, serta *weight* yang digunakan adalah *distance* karena akan menghitung jarak dengan yang paling dekat. Tahap berikutnya yaitu melakukan perbandingan algoritma klasifikasi dengan *test and score* untuk menghitung tingkat keberhasilan setiap algoritma klasifikasi di *orange* seperti gambar berikut ini:



Gambar 10 Hasil Test dan Score

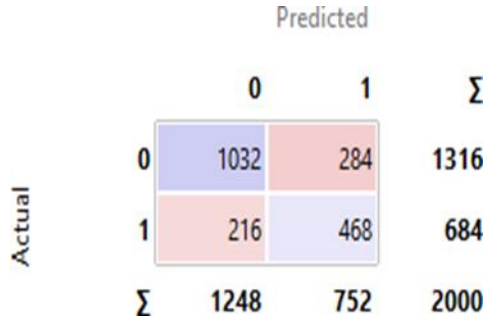
Menurut 2000 data yang sudah diuji, didapatkan hasil perhitungan *accuracy*, *Precision*, dan *recall* dari dari setiap algoritma seperti pada gambar IV.9. dari hasil klasifikasi algoritma *Naïve Bayes* dan K-*Nearest Neighbor* membuktikan bahwa hasil *accuracy* terbaik yaitu algoritma K-*Nearest Neighbor* yaitu sebanyak 99%.

juga bisa dilihat perbandingan dua algoritma AUC, ditemukan bahwa hasil nilai tertinggi AUC adalah algoritma K-*Nearest Neighbor* sejumlah 0.994. AUC digunakan untuk membandingkan satu model dengan model lainnya, semakin bagus nilai AUC maka semakin bagus hasil klasifikasi yang dipakai.

Selanjutnya dilakukan evaluasi menggunakan *confution matrix*, *confution matrix* digunakan untuk mengukur performa model klasifikasi dimana

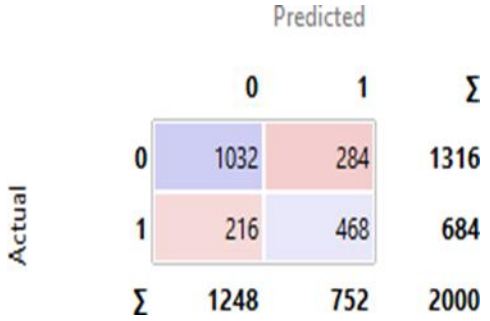
outpunya bisa berbentuk dua kelas atau lebih. *Confution matrix* adalah t*able* yang berisi nilai prediksi dan nilai *actual*. Hasil evaluasi setiap algoritma klasifikasi pada gambar berikut ini:

Tabel 1. Nilai Confusion Matrix Metode K-Nearest Neighbor



Dapat dijelaskan dari *confusion matrix* mengunakan algoritma K-*Nearest Neighbor* dapat dijelaskan bahwa Terdapat 2000 pasien sebanyak 1329 pasien terkena diabetes dan 671 pasien tidak terkena diabetes. Dari 1329 pasien yang terkena diabetes tersebut diprediksi benar sebanyak 1313 pasien dan diprediksi salah sebanyak 16 pasien. Kemudian Dari 671 pasien yang tidak terkena diabetes tersebut diprediksi benar sebanyak 668 pasien dan diprediksi salah sebanyak 3 pasien. Maka didapatkan hasi nilai accuracy 99%, Precision 98% dan Recall 99%.

Tabel 2. Nilai Confusion Matrix Metode Naives Bayes



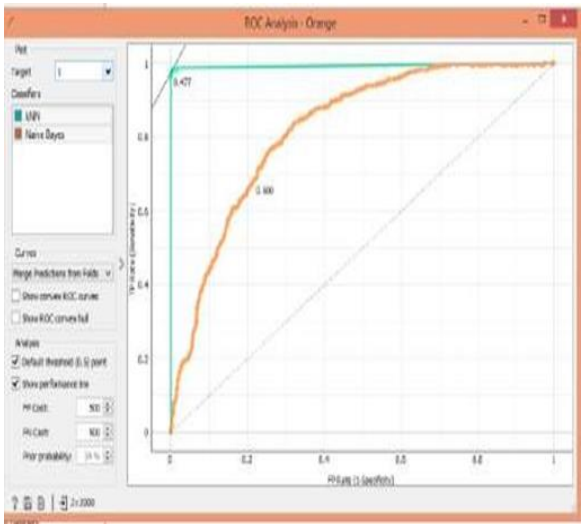
Dapat dijelaskan dari *confusion matrix* mengunakan algoritma *Naïve Bayes* dapat dijelaskan bahwaTerdapat 2000 pasien, sebanyak 1248 pasien terkena diabetes dan 752 pasien tidak diabetes. Dari 1284 pasien terkena diabetes tersebut diprediksi benar sebanyak 1032 pasien, dan diprediksi salah sebanyak 216 pasien. Kemudian dari 752 pasien yang tidak terkena diabetes tersebut diprediksi benar sebanyak 468 pasien, dan diprediksi salah sebanyak 284 pasien. Maka dihasilkan nilai accuracy 75%, Precision 82% dan Recall 78%.

Berdasarkan hasil evaluasi menggunakan model confusion matrix pada nilai evaluasi accuracy, prediction dan recall dari dua algoritma Naives Bayes dan K-Nearest Neighbor dapat dibandingkan pada tabel berikut:

Tabel 3. Hasil Perbandingan Evaluasi Algoritma

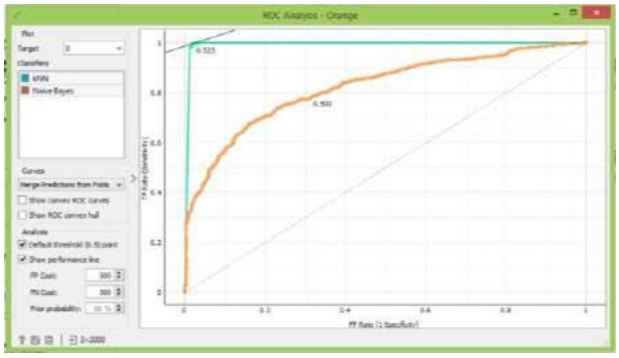
	Accuracy	*Precision*	*Recall*
Naïve Bayes	75%	82%	78%
K-*Nearest Neigbor*	99%	98%	99%

Diketahui bahwa kinerja dari model algoritma K-*Nearest Neighbor* lebih baik dari algoritma *Naïve Bayes* dengan nilai sebanyak 99%. Nilai *accuracy* juga bisa dilihat dengan menggunakan ROC *analysis* untuk melihat perbandingan yang divisualisasikan oleh *confusion matrix*. Melihat ROC *analysis* adalah cara yang mudah untuk melihat perbandingan nilai *accuracy* setiap algoritma klasifikasi secara grafik. Hasil grafik yang didapatkan dari ROC *analysis* dapat melihat pada gambar berikut ini:



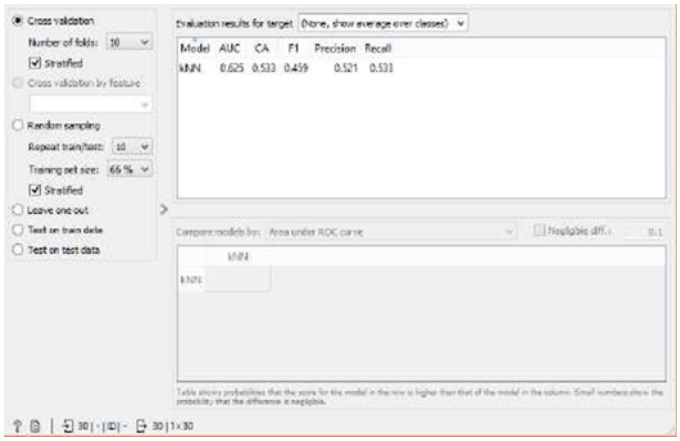
Gambar 11. ROC *Analysis* Prediksi Diabetes Dengan Target 0

Menunjukkan hasil ROC *analysis* hasil prediksi diabetes dengan target 0 pada setiap algoritma sebagai berikut K-*Nearest Neighbor* sejumlah 0.523 dan *Naïve Bayes* sejumlah 0.500 oleh karena itu untuk penelitian ini algoritma yang memiliki nilai *accuracy* paling baik adalah K-*Nearest Neighbor*.



Gambar 12. ROC *Analysis* Prediksi Diabetes Dengan Target 1

membuktikan hasil ROC *analysis* prediksi diabetes dengan target 1 pada setiap algoritma klasifikasi sebagai berikut K-*Nearest Neighbor* sejumlah 0.477 dan *Naïve Bayes* sejumlah 0.500.

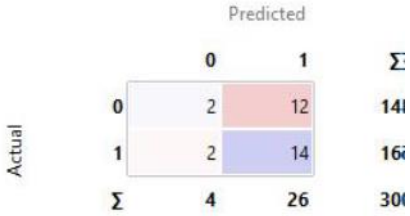


Gambar 13. Hasil Test and Score K-Nearest Neighbor

Berdasarkan 30 data yang sudah diuji, didapatkan hasil perhitungan *accuracy*, *Precision*, dan *recall* dari dari setiap algoritma seperti pada gambar 4.15. dari hasil klasifikasi algoritma K-*Nearest Neighbor* membuktikan bahwa hasil *accuracy* algoritma K-*Nearest Neighbor* yaitu sebanyak 70%. juga bisa dilihat bahwa hasil nilai AUC metode algoritma K-*Nearest Neighbor* sejumlah 0.625. AUC digunakan untuk membandingkan satu model dengan model lainnya, semakin besar hasil AUC maka semakin baik hasil klasifikasi yang dipakai.

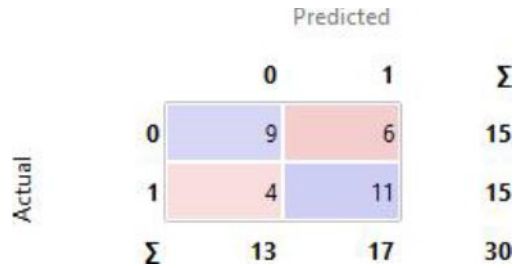
Selanjutnya dilakukan evaluasi menggunakan *confution matrix*, *confution matrix* digunakan untuk mengukur performa model klasifikasi dimana outputnya bisa berbentuk dua kelas atau lebih. Hasil evaluasi model K-*Nearest Neighbor* klasifikasi pada tabel berikut ini:

Tabel 4. Nilai Confusion Matrix Metode K-Nearest Neighbor



Dapat dijelaskan dari *confusion matrix* mengunakan algoritma K-*Nearest Neighbor* dapat dijelaskan berikut ini: Terdapat 30 data pasien sebanyak 4 pasien terkena diabetes dan 26 pasien tidak terkena diabetes. Diri 4 pasien yang terkena diabetes tersebut di prediksi benar sebanyak 2 pasien dan diprediksi salah sebanyak 2 pasien. Kemudian dari 26 pasien yang tidak terkena diabetes tersebut diprediksi benar sebanyak 14 dan di prediksi salah sebanyak 12 pasien.Maka dihasilkan nilai *accuracy 53%*, *precision 50%*, dan r*ecall 14%* dari algoritma K-*Nearest Neighbor* .

Tabel 5. Nilai Confusion Matrix Metode Naives Bayes



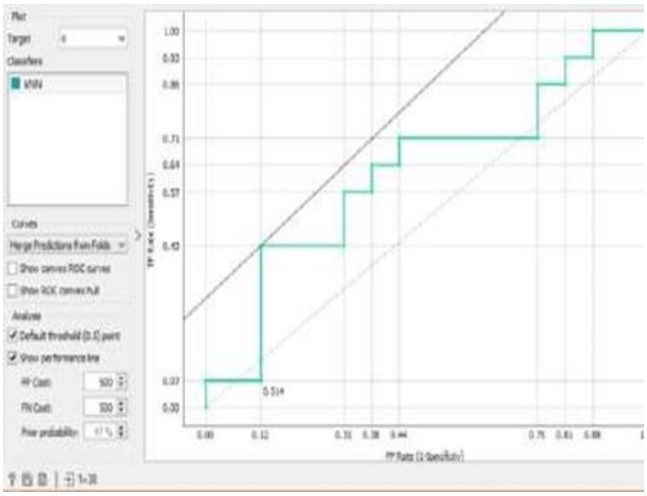
Dapat dijelaskan dari *confusion matrix* mengunakan algoritma *Naïve Bayes* dapat dijelaskan bahwaTerdapat 30 data pasien, sebanyak 13 pasien terkena diabetes dan 17 pasien tidak terkena diabetes. Dari 13 pasien yang terkena diabetes tersebut diprediksi benar sebanyak 9 pasien dan diprediksi salah sebanyak 4 pasien. Dari 17 pasien yang tidak terkena diabetes tersebut diprediksi benar sebanyak 11 pasien dan diprediksi salah sebanyak 6 pasien. Maka dihasilkan nilai *accuracy 66%*, *precision 69%*, dan *recall 60%* dari algoritma *Naïve Bayes.*

Berdasarkan hasil evaluasi menggunakan confusion matrix dihasilkan nilai perbandingan accuracy, prediction dan recall dari algoritma Naives Bayes dan K-Nearest Neightbor yang menggunakan 30 data disajikan pada tabel perbandingan berikut:

Tabel. 6. Hasil Perbandingan Evaluasi Algoritma

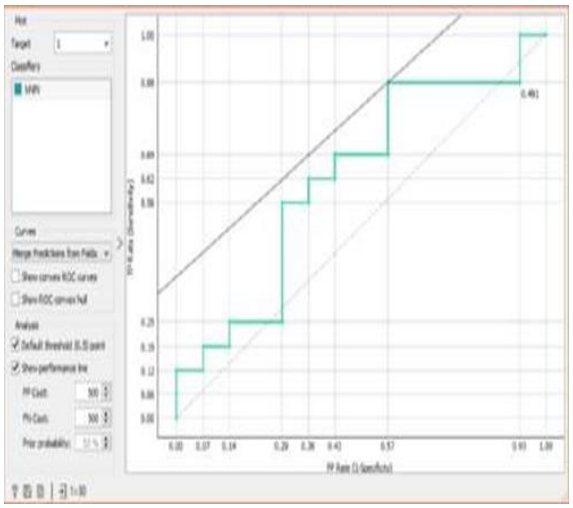
	Accuracy	*Precision*	*Recall*
Naïve Bayes	66%	69%	60%
K-*Nearest Neigbor*	53%	50%	14%

dapat diketahui bahwa performa dari algoritma *Naïve Bayes* lebih bagus dari algoritma K-*Nearest Neighbor* dengan hasil *accuracy* sebanyak 66%. Nilai *accuracy* dilihat dengan menggunakan ROC *analysis* untuk melihat perbandingan yang divisualisasikan oleh *confusion matrix*. Melihat ROC *analysis* adalah cara yang mudah untuk melihat perbandingan nilai *accuracy* setiap algoritma klasifikasi secara grafik. Hasil grafik yang didapatkan dari ROC *analysis* dapat melihat pada gambar berikut ini:



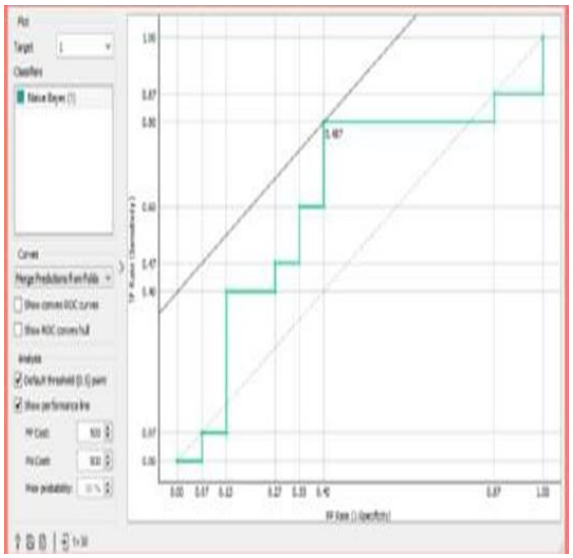
Gambar 14. *ROC* Analysis *Algoritma K-Nearest Neighbor Prediksi Diabetes Dengan Target 0*

Menunjukkan hasil ROC *analysis* hasil prediksi diabetes dengan target 0 pada model K-*Nearest Neighbor* adalah 0.514



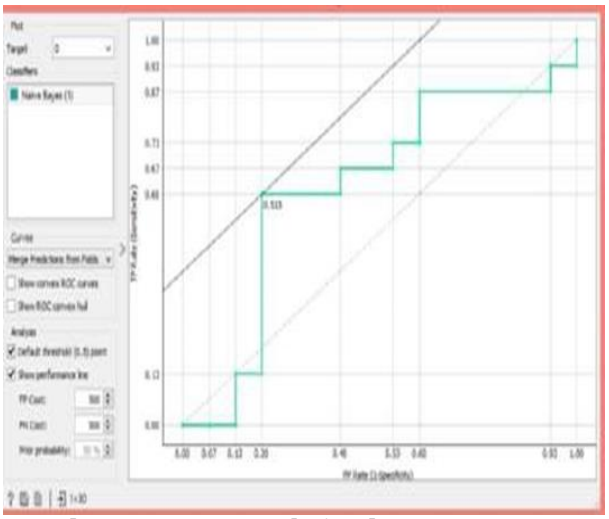
Gambar 15. *ROC* Analysis *Algoritma K-Nearest Neighbor Prediksi Diabetes Dengan Target* 1

Menunjukkan hasil ROC *analysis* hasil prediksi diabetes dengan target 0 pada model K-*Nearest Neighbor* adalah 0.491.



Gambar 16. ROC Analysis Algoritma Naïve Bayes Prediksi Diabetes Dengan Target 0

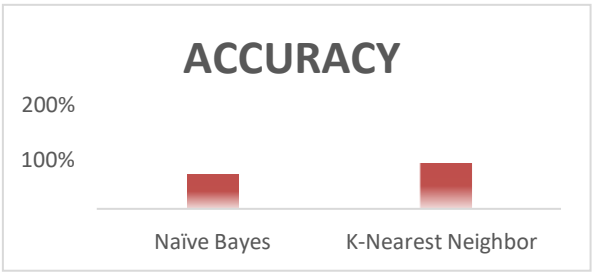
menunjukkan hasil ROC analysis hasil prediksi diabetes dengan target 0 pada algoritma Naïve Bayes adalah 0.515.



Gambar 17. ROC *Analysis* Algoritma *Naïve Bayes* Prediksi Diabetes Dengan Target 1

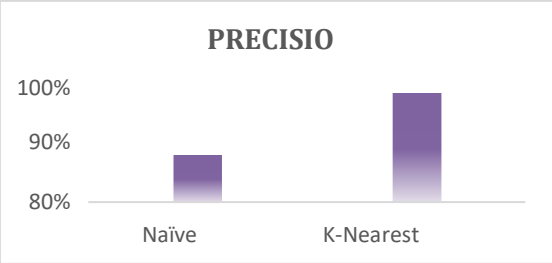
Menunjukkan hasil ROC *analysis* hasil prediksi diabetes dengan target 1 pada model *Naïve Bayes* adalah 0.487.

Hasil pengujian orange dengan memakai 10 – *fold cross validation* maka dapat diperoleh hasil *accuracy*, *precision*, dan *recall* yang bisa dilihat pada gambar berikut ini:



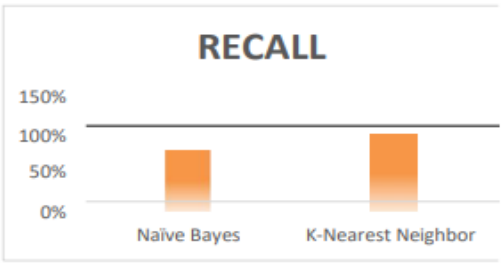
Gambar 18. Hasil Accuracy

dapat dijelaskan bahwa hasil nilai *accuracy* yang di dapat oleh algoritma *Naïve Bayes* sebanyak 75% dan *accuracy* algoritma K-*Nearest Neighbor* sebanyak 99%. Maka dapat disimpulkan hasil nilai *accuracy* terbaik adalah K-*Nearest Neighbor* dengan hasil *accuracy* sebanyak 99%.



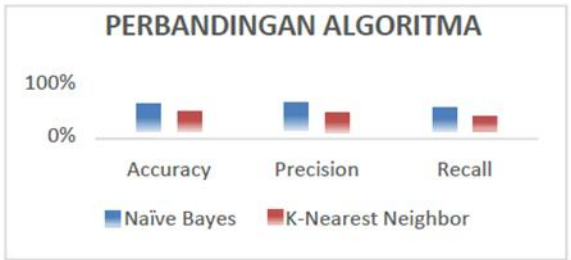
Gambar 19. Hasil Precision

Hasil nilai *precision* yang diperoleh dari algoritma *Naïve Bayes* sebanyak 82% dan algoritma K-*Nearest Neighbor* sebanyak 98%. Dapat disimpulkan bahwa nilai *precison* terbaik yaitu K- *Nearest Neighbor* dengan nilai *precision* sebanyak 98%.



Gambar 20. Hasil Recall

hasil *recall* yang didapat pada algoritma *Naïve Bayes* adalah 78% sedangkan algoritma K-*Nearest Neighbor* memperoleh hasil nilai 99%. Dapat disimpulkan bahwa hasil *recall* terbaik adalah algoritma K-*Nearest Neighbor* yaitu sebanyak 99%.



Gambar 21. Hasil Perbandingan Algoritma

Hasil nilai *accuracy* yang diperoleh yang menggunakan 30 data pada algoritma *Naïve Bayes* adalah 66% sedangkan algoritma K-*Nearest Neighbor* memperoleh hasil nilai 53%. Dapat disimpulkan bahwa hasil nilai *accuracy* terbaik menggunakan 30 data adalah algoritma *Naïve Bayes* yaitu sebanyak 66%.

Lalu hasil nilai *precision* yang diperoleh dari algoritma *Naïve Bayes* sebanyak 69% dan algoritma K-*Nearest Neighbor* sebanyak 50%.dapat disimpulkan bahwa hasil nilai *precision* terbaik adalah algoritma *Naïve Bayes* yaitu sebanyak 69%.

Dan hasil *recall* yang didapat pada algoritma *Naïve Bayes* sebanyak 60% sedangkan algoritma K- *Nearest Neighbor* memperoleh hasil nilai 14%. Dapat disimpulkan bahwa hasil *recall* terbaik adalah algoritma *Naïve Bayes* sebanyak 60%.

SIMPULAN

Berdasarkan dari hasil evaluasi dan analisis data mining klasifikasi pada dataset *Diabetes Prediction Using Logistic Regression* memakai algoritma *Naïve Bayes* dan K-*Nearest Neighbor* dengan menggunakan 10 – *fold cross validation.*

Maka dapat diberi kesimpulan bahwa dalam mengimplementasikan klasifikasi dataset *Diabetes Prediction Using Logistic Regression* untuk memprediksi diabetes dengan memakai algoritma *Naïve Bayes* dan K-*Nearest Neighbor* dari hasil *accuracy* yang didapat dengan memakai 10 – *fold cross validation* dengan 2000 data serta *Neighbor*- 13 untuk algoritma K-*Nearest Neighbor*, yang menghasilkan nilai *accuracy* terbaik adalah algoritma K-*Nearest Neighbor*. Sedangkan untuk melakukan uji coba model yang menggunakan 30 data serta *Neighbor*-18 untuk algoritma K-*Nearest Neighbor* yang menghasilkan nilai *accuracy* terbaik adalah algoritma *Naïve Bayes*.

Selanjutnya Dari hasil evaluasi yang menggunakan 2000 data K-*Nearest Neighbor* memiliki hasil *accuracy* sejumlah 99% sedangkan *Naïve Bayes* memiliki hasil *accuracy* sejumlah 75% selisih akurasi terhadap dua metode tersebut sejumlah 24%. Dari hasil evaluasi uji coba model mengunakan 30 data yang dibagi menjadi

data uji dan data latih K-*Nearest Neighbor* memiliki hasil *accuracy* sejumlah 53% dan untuk *Naïve Bayes* memiliki hasil a*ccuracy* sejumlah 66% meiliki selisih *accuracy* sejumlah 13 %.

Maka dapat disimpulkan bahwa perbandingan implementasi untuk mengimplementasikan prediksi diabetes dengan menggunakan 2000 data dan 30 data untuk uji coba model K-*Nearest Neighbor* dan *Naïve Bayes* menghasilkan *accuracy* yang berbeda. Dalam implementasi untuk mengklasifikasikan prediksi diabetes dengan 2000 data algoritma K-*Nearest Neighbor* dapat menghasilkan tingkat nilai *accuracy* prediksi diabetes lebih baik dibandingkan dengan memakai algoritma *Naïve Bayes*. sedangkan yang menggunakan data untuk uji coba model yaitu 30 data algoritma *Naïve Bayes* dapat menghasilkan tingkat nilai *accuracy* prediksi diabetes lebihbaik dibandingkan dengan memakai alogoritma K- *Nearest Neighbor* . Berdasarkan hasil evaluasi tersebut algoritma KNN lebih mendukung pencarian data dalam jumlah besar sehingga lebih sesuai untuk penggalian dataset diabetes dengan jumlah record 2000 data.

DAFTAR PUSTAKA

Argina, A. (2020). Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes. *Indonesian Journal od Data and Science*, 29-33.

Argina, A. M. (2020). Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes. *Indonesia Jurnal Data Science*, 29–33.

B., M. P. (2020). Komparasi Algoritma KNN dan Naives Bayes untuk Klasifiksi Diagnosis Penyakit Diabetes Mellitus. *Journal Sain dan Manajemen*, 45-57.

Bingga, I. A. (2021). Kaitan Kualitas Tidur dengan Diabetes Melitus Tipe 2. *Med. Hutama*, 1047–1052.

Connely, L. (2020). Logistic Regression. *Medsurg Nursing: Pitman*, 353.

F., K. P. (2018). Klasifikasi Diabetes Menggunakan Model Pembelajaran Ensemble Blending, *ULTIMATICS*, pp. 11-15.

Hendrik, N. S. (2018). Komparasi Kinerja Algoritma Data Mining pada Dataset Konsmsi Alkohol Siswa,". *Khazanah Informatika*, pp. 98-103.

Hozairi Hozairi, A. A. (2021). Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes. *Jurnal Ilmiah Network Engineering Research Operation*, 133-144.

Joshua Neumiller, G. B. (2020). 1. Improving care and promoting health in populations: Standards of medical care in diabetes- 2020. *Diabetes Care*, S7–S13.

N. Maulidah, R. S. (2021). "Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes. *Indonesian Journal of Software Engineering*, pp. 63–68.

Novrizal Eka Saputra, K. D. (2016). Penerapan Knowledge Management System (KMS) Menggunakan Teknik Knowledge Data Discovery (KDD) pada PT. PLN (Persero) WS2JB Rayon Kayu Agung. *Jurnal Sistem Informasi*, pp. 1038-1055.

Nurdiana, N., Rodiyansyah, S. F., & Algifari, A. (2020). Studi Komparasi Algoritma ID3 dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. *INFOTECH journal*, *6*(2), 18-23.

Patwari, B. A. (2021). "Komparasi Metode Klasifikasi Data Mining Decission Tree dan Naives Bayes untuk Prediksi Penyakit Diabetes,". *Infotek Jurnal Informatika Dan Teknologi*, pp. 63-69.

Permadi. J., R. H. (2021). " Perbandingan K-Nearest Neigtbor dan Backpropagation Neural Network dalam Prediksi Resiko Penyakit Diabetes Tahap Awal. *Jurnal Ilmu Komputer*, pp. 352-365.

Qatrunnada Refa Cahyani, M. J. (2022). Prediksi Resiko Penyakit Diabetes Menggunakan Algoritma Regressi Logistik. *Journal of Machine Learning and Artificial Intelligence*, pp.107-114.

Rahayu P.T., D. d. (2022). Perbandingan Algoritma K-Nearest Neighbor dan Gaussian Naives Bayes pada

klasifikasi penyakit diabetes melitus. *Jurnal Smart Teknologi*, pp. 366-373.

Ridwan, A. (2020). Penerapan Algoritma Naïve Bayes Untuk KlasifikasiPenyakit Diabetes Mellitus. *Jurnal Sistem Komputer dan Kecerdasan Buatan*, pp 15-21.

Wibawa, M. S. (2018). Prediksi Penyakit Diabetes Menggunakan Algoritma ID3 dengan Pemilihan Atribut Terbaik (Diabetes Prediction using ID3 Algorithm with Best Attribute Selection). *Juita*, pp. 29–35.