Jurnal Komtika (Komputasi dan Informatika),
Vol. 5 No. 1| Mei 2021

Komparasi Algoritma *Naïve Bayes* dan *K-Nearest Neighbor* untuk Membangun Pengetahuan Diagnosa Penyakit Diabetes

Maulidya Dwi Nurmalasari[1*], Kusrini[2], Sudarmawan[3]
[1,2,3] Magister Teknik Informatika , Universitas Amikom Yogyakarta
*email: maulidya.302@students.amikom.ac.id

DOI:https://doi.org/10.31603/komtika.v5i1.5140

Received: 10-06-2021, Revised: 29-06- 2021, Accepted:03-07- 2021

ABSTRACT

Diabetes is caused by a deficiency of the hormone insulin, which is secreted by the pancreas to lower blood sugar levels. The factors that trigger the occurrence of diabetes are derived from various factors such as a combination of genetic and environmental factors. The phenomenon of the emergence of various beverage brand outlets can be one of the triggers for blood sugar levels in humans. Normal blood sugar levels in the body range from 70-130 mg/dL before eating, less than 180 mg/dL two hours after eating, less than 100 mg/dL after not eating or surviving for eight hours, and 100-140 mg/dL at bedtime. This research aims to determine which algorithm is suitable for building knowledge about diabetes using the Naïve Bayes and K-Nearest Neighbor (KNN) algorithm. The accuracy results from Naïve Bayes are 85.60% and K- Nearest Neighbor of 91.61%. The results showed that K-Nearest Neighbor proved to have the best accuracy.

Keywords: *K-Nearest Neighbor(KNN) , Naïve Bayes, Diabetes, Comparison*

ABSTRAK

Diabetes disebabkan oleh kekurangan hormon insulin yang dikeluarkan oleh pankreas untuk menurunkan kadar gula darah. Faktor-Faktor yang memicu terjadinya penyakit diabetes berasal dari berbagai faktor seperti kombinasi faktor genetik dan lingkungan.Munculnya fenomena munculnya berbagai outlet brand minuman bisa menjadi salah satu pemicu kadar gula darah pada manusia. Kadar gula darah normal pada tubuh berkisar antara 70-130 mg/dL pada saat sebelum makan, kurang dari 180 mg/dL pada saat dua jam setelah makan, kurang dari 100 mg/dL pada saat setelah tidak makan atau berpuasa selama delapan jam, dan 100-140 mg/dL pada saat menjelang tidur. Penelitian ini bertujuan untuk mengetahui algoritma mana yang cocok untuk membangun pengetahuan penyakit diabetes menggunakan algoritma *Naïve Bayes* dan *K- Nearest Neighbor* (KNN). Hasil akurasi dari *Naïve Bayes* yaitu 85,60% dan *K- Nearest Neighbor* sebesar 91,61%. Hasil penelitian menunjukkan bahwa terbukti algoritma *K- Nearest Neighbor* sebagai algoritma yang memiliki akurasi terbaik.

Kata-kata kunci: *K-Nearest Neighbor(KNN) , Naïve Bayes*, Diabetes, Komparasi

PENDAHULUAN

Diabetes merupakan salah satu penyakit yang sangat menakutkan bagi sebagian besar orang didunia. Diabetes diklasifikasikan menjadi 2 jenis yaitu diabetes tipe 1 dan tipe 2.[1] Menurut Kementrian Kesehatan Republik Indonesia (Kemenkes), diabetes disebabkan oleh kekurangan hormon insulin yang dikeluarkan oleh pankreas untuk menurunkan kadar gula darah. Kombinasi faktor genetik dan lingkungan juga memicu terjadinya diabetes mellitus type 2. Kadar gula darah normal pada tubuh berkisar antara 70-130 mg/dL pada saat sebelum makan, kurang dari 180 mg/dL pada saat dua jam setelah makan, kurang dari 100 mg/dL pada saat setelah tidak makan atau berpuasa selama delapan jam, dan 100-140 mg/dL pada saat menjelang tidur [2]. Kemenkes juga menyatakan ada beberapa faktor resiko penyakit diabetes

yang bisa diubah diantaranya adalah kegemukan (Berat badan lebih /IMT > 23 kg/m2) dan lingkar perut (Pria > 90 cm dan Perempuan > 80cm), kurang olahraga/aktivitas fisik, dislipidemia (Kolesterol HDL ≤ 35 mg/dl, trigliserida ≥250 mg/dl, riwayat penyakit jantung, hipertensi/ tekanan darah tinggi (> 140/90 mmHg) dan diet tidak seimbang. Diabetes tidak hanya disebabkan oleh kadar gula yang tinggi pada darah, ternyata juga ada faktor lain yang meningkatkan resiko penyakit diabetes. Faktor lain inilah yang perlu diperhatikan. Jika menggunakan sistem pakar maka pakar juga tidak akan sepenuhnya yakin berapa besar nilai faktor dari setiap faktor yang mempengaruhi. Ketidakyakinan pakar ini tentu akan menurunkan akurasi jika menggunakan sistem pakar sebagai pembangun pengetahuan mengenai diabetes.

Beberapa pakar melakukan penelitian untuk mendiagnosa penyakit diabetes. Salah satu contoh pakar mengimplementasikan sistem pakar deteksi penyakit diabetes mellitus (DM) dengan menggunakan metode *forward chaining* dan *certainty factor* yang dimana tidak tercantum asal data gejala yang didapat dari pakar atau dengan observasi dan gejala yang masih sedikit untuk mendeteksi penyakit diabetes. Peneliti selanjutnya juga melakukan penelitian tentang diagnosa penyakit diabetes dengan metode *forward chaining* dengan terdapat banyak gejala yang tidak bisa dipastikan memang gejala dari penyakit diabetes.[3],[4]

Dalam penelitiannya [5] melakukan komparasi terhadap kejang pada penyakit epilepsi dengan tes EGG menggunakan 3 metode yaitu *Naïve Baues, Random Tree Forest* dan *K-Nearest Neighbor*(KNN) dengan akurasi terbaik adalah KNN 92,7%, *random tree forest* 86,6%) dan *Naïve Bayes* sebesar 55,6% . Penelitian pada penyakit thyroid dengan menggunakan beberapa algoritma yaitu *Naïve Bayes* dengan akurasi 85% , *support vector machine* (SVM) dengan akurasi 82%, dan KNN dengan akurasi sebesar 85% telah dilakukan [6]. Selanjutnya penelitian pada penyakit diabetes mellitus dengan menggunakan beberapa algoritma *machine learning* yaitu KNN dengan akurasi 100% , *Naïve Bayes* dengan akurasi 86%, *Logistic Regression* dengan akuarsi76%, *Random Forest Classifier* dengan akurasi 82%, *Decision Tree Classifier* dengan akurasi 80%, dan SVM dengan akurasi 80% juga telah dilakukan [7].

Selain itu, penelitian yang dilakukan dengan sistem pakar yang terdapat pada gejala penyakit diabetes masih belum lengkap dan ada yang menggunakan gejala yang banyak tetapi tidak terbukti sumbernya dari mana sehingga diharapkan dengan adanya penelitian ini untuk pembentukan pengetahuan diagnosis penyakit diabetes dapat membantu para pakar dalam menentukan pengetahuan tentang penyakit diabetes. Penelitian ini membangun pengetahuan melalui klasifikasi dengan algoritma *Naïve Bayes* dan KNN sehingga dari hasil klasifikasi yang berasal dari history penderita diabetes maka data history penderita penyakit diabetes dapat dijadikan untuk membangun pengetahuan agar nantinya dapat membantu pakar dalam menentukan gejala dari diabetes. Algoritma *Naïve Bayes* menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari data set yang diberikan sedangkan algoritma KNN membandingkan jarak kedekatan antara data *training* dan data *testing* [8],[9].

Penelitian ini melakukan komparasi algoritma *Naïve Bayes* dan KNN untuk membangun pengetahuan tentang diagnosa penyakit diabetes dengan menghitung nilai perfoma kedua model algoritma tersebut. Ada tambahan perhitungan yang digunakan yaitu *F1-Score* yang memiliki fungsi untuk memahami algoritma mana yang lebih sesuai dengan data berdasarkan nilai *Precision* dan *Recall.*

METODE

Pada penelitian ini menggunakan metode klasifikasi dalam membangun pengetahuan diagnosa penyakit diabetes menggunakan algoritma *Naïve Bayes* dan KNN dengan menerapkan perhitungan perfoma melalui perhitungan *confussion matrix* yaitu *accuracy*, *precision*, *recall,* dan *f1- score.* Data set yang akan digunakan pada penelitian ini dari UCI *Machine Learning Repository*.

Klasifikasi merupakan langkah atau cara dalam upaya membentuk suatu model atau fungsi yang digunakan dalam menjelaskan atau membedakan konsep kelas data[10]. Proses klasifikasi dilakukan terbagi dengan dua tahapan yaitu pelatihan dan pengujian yang bertujuan agar komputer dapat belajar mengenal beberapa objek(data latih). *Naive Bayes* dan KNN termasuk dalam klasifikasi [11]. *Naïve Bayes* merupakan sebuah metode penggolongan berdasarkan probabilitas sederhana dan dirancang untuk dipergunakan dengan asumsi bahwa antar satu kelas dengan kelas yang lain tidak saling tergantung . Formulasi *Naïve Bayes* dinyatakan dalam persamaan 1[12].

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad (1)$$

dimana:

$P(Y|X)$ adalah probabilitas data dengan vector X pada kelas Y

$P(Y)$ adalah probabilitas awal kelas Y (*prior probability*)

$\prod_{i=1}^{q} P(X_i|Y)$ adalah probabilitas independen kelas Y dari semua fitur dala vector X

Nilai $P(X)$ adalah probabilitas dari X

Klasifikasi K-NN merupakan algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain. Dekat atau jauhnya lokasi (jarak) biasanya dihitung berdasarkan jarak *Euclidean* dengan rumus seperti pada persamaan (2):

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{N} (\text{diff}\,(x_{il}, x_{jl}))^2} \quad (2)$$

dengan :

x_{il} adalah data *testing* ke-i pada variabel ke-1

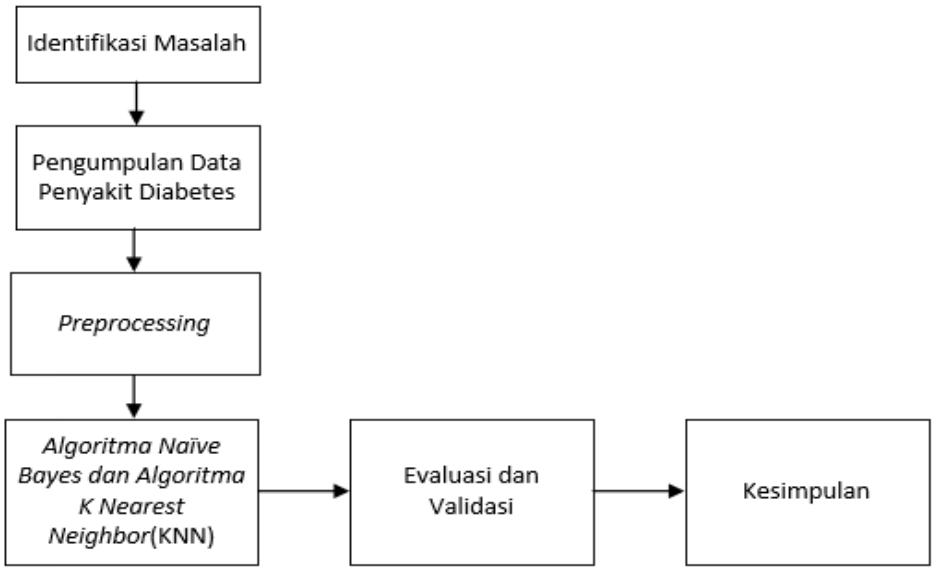
x_{ij} adalah data *training* ke-i pada variabel ke-l

$d(x_i, x_j)$ adalah jarak

N adalah dimensi data variabel bebas

$\text{diff}(x_{il}, x_{jl})$ adalah *difference* atau ketidaksamaan

Tahapan dalam penelitian ini terdiri dari beberapa tahapan seperti pada Gambar 1.



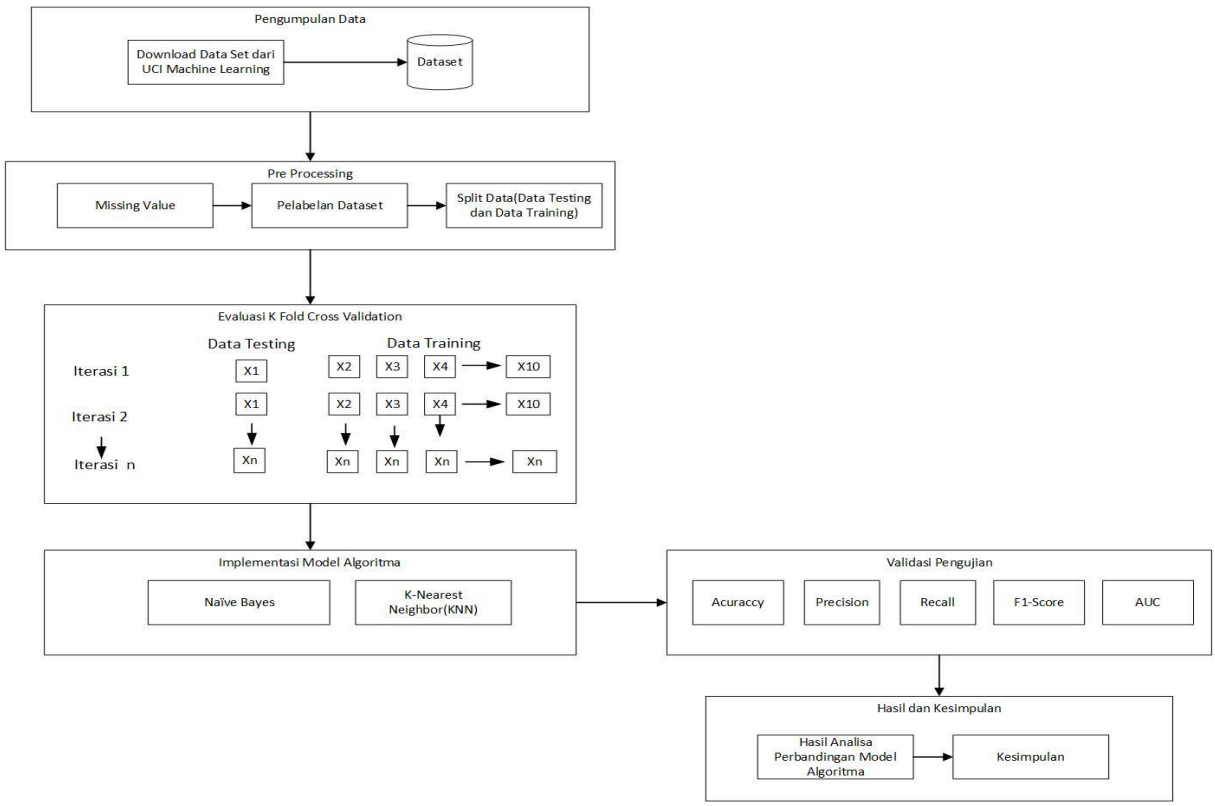
Gambar 1. Alur Penelitian

Dalam tahap identifikasi masalah dilakukan identifikasi masalah dengan mencari permasalahan yang ada pada objek penelitian. Tahap berikutnya adalah proses pengumpulan data berupa history dari diagnosis penyakit diabetes dari sumber internet yaitu dari UCI *Machine Learning Repository. Pada tahap preprocessing* dilakukan pengolahan terhadap kumpulan history diagnosis penyakit diabetes dari UCI Machine Learning Repository sebelum dilakukan klasifikasi. Metode klasifikasi dilakukan menggunakan algoritma *Naïve Bayes dan* KNN. Pembobotan *Naïve Bayes* dilakukan melalui probabilitas yang diambil dari dataset dan untuk KNN dilakukan dengan konfigurasi nilai K = 3. Proses perhitungan klasifikasi dilakukan dengan menggunakan tools Rapid Miner.

Selanjutnya dilakukan tahap evaluasi dan validasi pada model klasifikasi yang telah dibuat untuk setiap skenario.Validasi pada penelitian ini menggunakan 10 k–fold lalu selanjutnya melakukan evaluasi yang didapatkan dari nilai-nilai *confusion matrix* dari setiap skenario untuk mendapatkan nilai *accuracy, precision*, *recall*, dan *F1- score* dari model klasifikasi dengan hasil data dalam bentuk kurva *Receiver Operating Characteristic* (ROC) untuk mengukur nilai *Area Under Curve* (AUC). Tahap terakhir adalah kesimpulan dengan menyajikan hasil dari percobaan yang telah dilakukan dengan beberapa fakta terkait arsitektur *Naïve Bayes* dan KNN terhadap tingkat akurasi yang memiliki kinerja terbaik.

HASIL DAN PEMBAHASAN

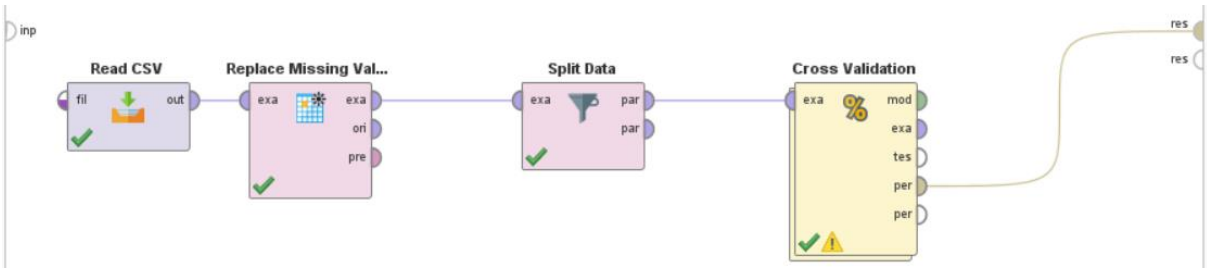
Hasil yang didapatkan dalam melakukan penelitian ini meliputi beberapa tahapan seperti pada Gambar 2.



Gambar 2. Detail Tahapan Penelitian

Pengumpulan data penyakit diabetes dilakukan dengan menggunakan *dataset* dari UCI *Machine Learning Repository* yang didownload dari https://archive.ics.uci.edu/ml/machine-learning-databases/00529/. Dataset penyakit diabetes terdiri dari 521 data dengan 17 feature atau atribut yaitu *Age, Gender, Polyuria, Polydipsia, Sudden Weight Loss, Weakness,Polyphagia, Genital Thrush, Visual Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiffness, Alopecia, Obesity,* dan *Class*.

Pada tahap *preprocessing* dilakukan menggunakan Rapid Miner dengan memberikan label kolom yang akan diklasifikasi dan menghilangkan data yang kosong seperti disajikan dalam Gambar 3. Proses *prepocessing* dilakukan menggunakan *Replace Missing Value* untuk menghilangkan data yang berulang. Setelah melakukan tahap *preprocessing* selanjutnya melakukan *split* data yaitu membagi data dengan rasio 80% untuk data *training* dan 20 % sebagai data *testing* yang akan digunakan untuk proses klasifikasi. Terakhir dilakukan *cross validation* untuk validasi data 10 *k-fold* sebelum data akan diklasifikasikan.

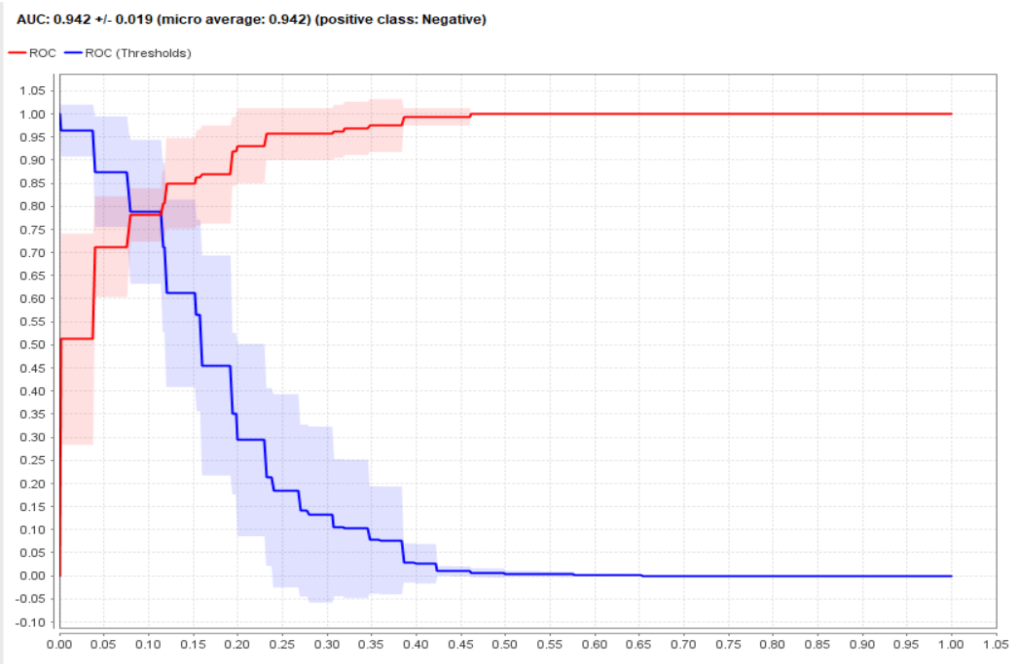


Gambar 3. *Preprocessing* Data

Setelah melakukan *preprocessing*, tahap selanjutnya adalah implementasi algoritma Naïve Bayes dan KNN. Untuk implementasi algoritma KNN dilakukan dengan konfigurasi nilai K=3. Setelah model klasifikasi menggunakan *Naïve Bayes* dan KNN dilakukan maka selanjutnya dilakukan pengujian model dan diperoleh nilai akurasi yang dinyatakan dalam *confussion matrix.* Hasil dari perhitungan perfoma untuk algoritma *Naïve Bayes* dan KNN dengan menggunakan validasi data yaitu 10 *k-fold* yang telah dilakukan pada saat *preprocessing* dan perhitungan perfoma dengan *confussion matrix* yaitu *accuracy*, *precision*, *recall*, *f1- score (f measure)* disajikan pada Tabel 1. Hasil dari evaluasi model klasifikasi dinyatakan dengan AUC untuk Algoritma *Naïve Bayes* dan KNN disajikan dalam Gambar 4 dan Gambar 5.

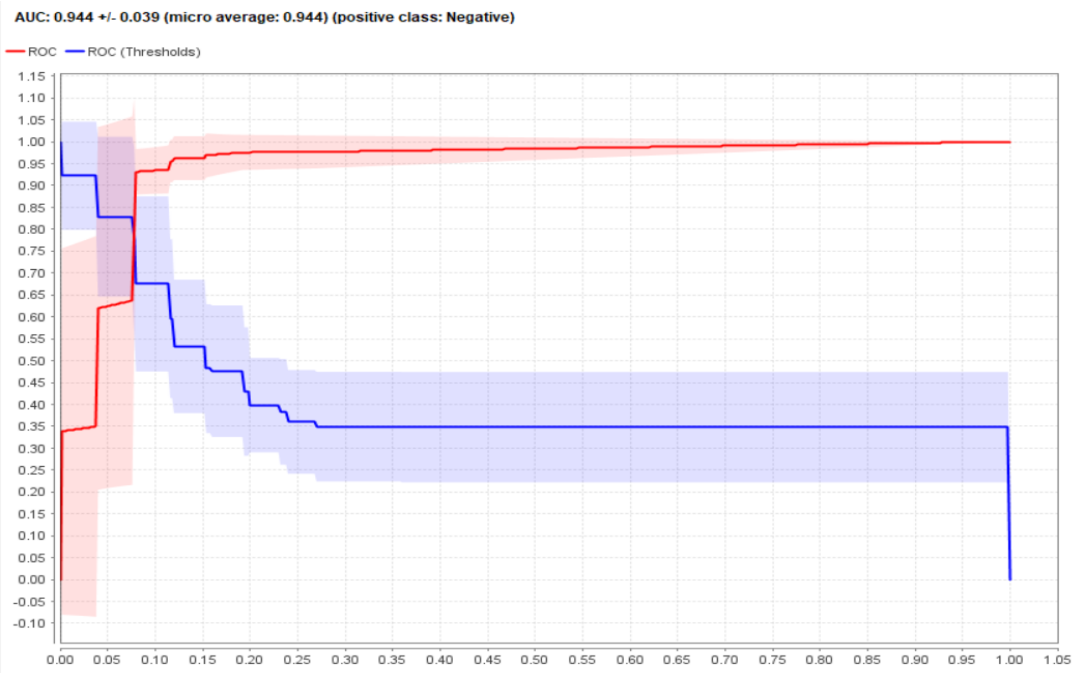
Tabel 1. Hasil Perhitungan Perfoma

Klasifikasi	*Accuracy*	*Precision*	*Recall*	*F Measure*
Naïve Bayes	85,60%	78,26%	88,12%	82,51%
K-Nearest Neighbor(KNN)	91.61%	86.38%	93.75%	89.66%



Gambar 4. Hasil AUC *Naïve Bayes*

Dari Gambar 4 dijelaskan bahwa hasil dari AUC *Naïve Bayes* yaitu kurva ROC yang digunakan untuk memperlihatkan data dari *confussion matrix*. Garis *horizontal* yang berarti nilai *False Positive* (FP) dan garis *vertical* yang berarti nilai *True Positive* (TP). Hasil AUC *Naïve Bayes* yaitu 0,942 yang menunjukan bahwa algoritma *Naïve Bayes* merupakan klasifikasi yang baik.



Gambar 5. Hasil KNN

Dari Gambar 5 diperoleh bahwa hasil dari AUC KNN yaitu kurva ROC yang digunakan untuk memperlihatkan data dari *confussion matrix*. Garis horizontal yang berarti nilai FP dan garis vertikal yang berarti nilai TP. Hasil AUC KNN pada Gambar 5 yaitu 0,944 yang menunjukan bahwa algoritma KNN merupakan klasifikasi yang baik.

KESIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa algorima KNN mempunyai performa lebih baik dibandingkan *Naïve Bayes* karena memiliki nilai *confusion matrix* yang tinggi dalam aspek *accuracy*, *precision*, *recall*, dan *f measure.* Oleh karena itu algoritma KNN dapat digunakan untuk membangun pengetahuan diagnosa penyakit diabetes agar dapat membantu pakar untuk mendiagnosa penyakit diabetes.

DAFTAR PUSTAKA

[1] Ramesh D and Katheria Y S (2019) Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach Health Technol. (Berl)

[2] Putra, S. (2019, November 27). Kadar Gula darah Normal Berapa Sih?. Diakses pada 6 Mei 2021. https://health.detik.com/berita-detikhealth/d-4801052/ kadar-gula-darah-normal-berapa-sih

[3] Yulianti et all (2021) Sistem Pakar Deteksi Penyakit Diabetes Mellitus (DM) menggunakan Metode Forward chaining dan Certainty factor Berbasis Android , Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi) 5 (1) 2021, 49-55

[4] S. Hardani (2020), “Diagnosa Penyakit Diabetes Dengan Metode Forward Chaining”, jitk(Jurnal Ilmu Pengetahuan dan Teknolog), vol. 5, no. 2, pp. 231-236, Feb. 2020.

[5] Fauzia et all (2019), "Epileptic Seizure Detection in EEGs by Using Random Tree Forest, Naïve Bayes and KNN Classification"Journal of Physics: Conference Series. 1505 012055

[6] Dr. Dayanand Jamkhandikar , Neethi Priya, Johan(2020),"Thyroid Disease Prediction Using Feature Selection And Machine Learning Classifiers", The International journal of analytical and experimental modal analysis. ISSN NO:0886-9367

[7] A. R. Bindiya , K. Nikhil , M. S. Sindhu Rashmi, Shafinaz Banu (2020), "Diabetes Mellitus Prediction using Machine Learning Algorithms",International Journal for Research in Applied Science & Engineering Technology (IJRASET). ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VII July 2020

[8] Patil, T. R., Sherekar, M. S., (2013), Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal of Computer Science and Applications, Vol. 6, No. 2, Hal 256-261.

[9] Asahar Johar T, Delfi Yanosma, Kurnia Anggriani. 2016, Implementasi Metode K-Nearest Neighbor (Knn) Dan Simple Additive Weighting (Saw) Dalam Pengambilan Keputusan Seleksi Penerimaan Anggota Paskibraka, Jurnal Pseudocode, Volume III Nomor 2, September 2016, ISSN 2355-5920

[10] Santoso, B. 2007. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu

[11] Han, J and Kamber, M. 2006. Data Mining Concepts and Techniques, second edition. California: Morgan Kaufman.

[12] Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: ANDI Yogyakarta