Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Perbandingan Algoritma Naïve Bayes dan *K-Nearest Neighbor (KNN)* untuk Mengetahui Keakuratan Diagnosa Penyakit Diabetes

Qonitah Alia Puteri [a], *Tri Sagirani* [b*], *Julianto Lemantara* [c]

[a,b,c] *Sistem Informasi, Universitas Dinamika, Jawa Timur*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 01 November 2023
Revisi Akhir: 28 Desember 2023
Diterbitkan *Online*: 31 Desember 2023

KATA KUNCI

Diabetes,

Knowledge Discovery in Databases (KDD),

Data Mining

Naïve Bayes,

K-Nearest Neighbor (KNN).

KORESPONDENSI

E-mail: tris@dinamika.ac.id*

A B S T R A C T

Diabetes adalah gangguan metabolisme kronis di mana pankreas tidak bisa menciptakan insulin yang cukup ataupun tubuh tidak berhasil menggunakan insulin yang telah dihasilkan. Penyebab yang dapat menimbulkan terjadinya penyakit diabetes berawal dari banyak antara lain penyebab genetik dan lingkungan. Diabetes tidak hanya menyebabkan kerugian ekonomi bagi pasien karena biaya pengobatan tetapi juga memperpendek umur peluang untuk hidup sebesar 5 – 10 tahun. Akibat lainnya adalah jika tidak ada upaya untuk mengontrol dan mencegah, diabetes dapat semakin memperburuk penderita karena dapat menimbulkan komplikasi yang berat. Berdasarkan permasalahan tersebut dapat dilakukan prediksi penyakit diabetes untuk dapat membantu tenaga medis mengetahu lebih dini kondisi pasien. Algoritma Naive Bayes dan K-Nearest neighbor (KNN) bisa digunakan membantu prediksi penyakit diabetes dengan menggunakan software RapidMiner & Python. Hasil penelitian ini dievaluasi dengan Confusion Matrix serta Nilai AUC. Hasil metode Naïve Bayes adalah 77% dengan nilai AUC 0.83 sedangkan metode K-nearest neighbor (KNN=3) adalah 71% dengan nilai AUC 0.75, KNN=5 adalah 69% dengan AUC 0.76, dan KNN=7 adalah 68% dengan AUC 0.75 sehingga dapat disimpulkan bahwa algoritma naïve bayes lebih unggul dibandingkan dengan KNN, meski pada penelitian ini untuk algoritma KNN menggunakan K=3, K=5 dan K=7, lalu untuk untuk yang KNN dari ketiga K yang digunakan dari segi *confusion matrix* KNN=3 lebih unggul sedangkan dari nilai AUC yaitu KNN=5.

1. PENDAHULUAN

Diabetes adalah sebuah penyakit yang kerap kali sering dijumpai oleh kita semua, penyakit diabetes juga sulit dideteksi dikarenakan gejala yang terjadi tidak seperti orang yang terkena penyakit. Terdapat beberapa tanda seseorang mengidap penyakit diabetes yaitu hiperglikemia, Hiperglikemia adalah kadar glukosa di dalam darah meningkat melampaui batas normal. Menurut Organisasi International Diabetes Federation (IDF) memprediksi bahwa ditemukan 463 juta orang di usia 20 -79 di dunia mengalami penyakit diabetes, berdasarkan jenis kelamin, IDF mengungkapkan bahwa pada 2019 sekitar 9% pada perempuan dan 9,65% pada laki – laki. Proporsi pada penyakit diabetes bertambah dengan seiringnya usia masyarakat naik hingga 111,2 juta orang di usia 65 – 79 tahun.

Diabetes juga adalah sebab yang utama terjadinya penyakit ginjal, kebutaan pada umur 65 tahun, dan juga amputasi [1]. Diabetes tidak hanya menyebabkan kerugian ekonomi bagi pasien karena biaya pengobatan tetapi juga memperpendek umur peluang untuk hidup sebesar 5 – 10 tahun [2]. Akibat lainnya adalah jika tidak ada upaya untuk mengontrol dan mencegah, diabetes dapat semakin memperburuk penderita karena dapat menimbulkan komplikasi yang berat [3].

Diagnosis medis diabetes sendiri cukup menghadapi rintangan, bahkan telah terjadi penurunan data. Data medis dengan banyak karakteristik yang tidak selaras dan berlebihan dapat

https://doi.org/10.25077/TEKNOSI.v9i3.2023.247-254

mempengaruhi bobot diagnosis penyakit [4]. Beberapa penelitian sudah pernah dilakukan sebelumnya untuk mendiagnosa penyakit diabetes. Salah satunya ialah penelitian tersebut menggunakan dataset yang terdiri dari 17 atribut antara lain, *Age, Gender, Polyuria, Polydipsia, Sudden Weight Loss,dll* [5]. Pentingnya untuk mengembangkan metode yang lebih efektif untuk mendeteksi, mendiagnosis dan mengobati diabetes.

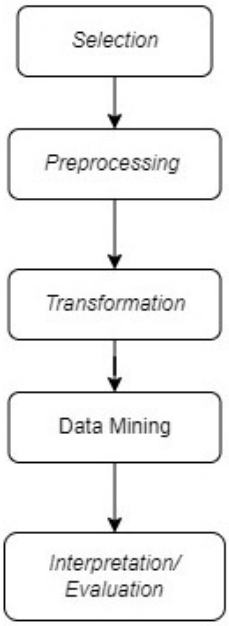
Dalam hal ini, penggunaan teknik data mining khususnya yang berkaitan dengan klasifikasi, telah menjadi alat yang cukup banyak digunakan untuk memahami dan mendiagnosis gejala diabetes [6]. Data mining adalah pendekatan analisis data yang mengungkap pola tersembunyi dan wawasan berharga dari data yang besar dan kompleks [7]. Dalam konteks diabetes, pengumpulan data dapat digunakan untuk mengidentifikasi faktor risiko, memprediksi perkembangan penyakit, dan mendukung pengambilan keputusan klinis [8].

Dalam penelitian klasifikasi pada penyakit diabetes mellitus dengan menerapkan Metode K-Nearest neighbor (KNN) mendapatkan Accuracy 96%. Penelitian pada komparasi diagnosis penyakit diabetes menerapkan 2 Algoritma yaitu *Correlated Naïve Bayes* dan Naïve Bayes dengan *Accuracy* yang paling baik adalah Correlated Naïve Bayes 67,15%, Naïve Bayes 64,33% [9]. Penelitian klasifikasi pada penyakit diabetes mellitus dengan menerapkan Metode K-Nearest neighbor (KNN) dengan hasil *accuracy* sebesar 96% [10]. Lalu pada penelitian perbandingan hasil analisis teknik data mining untuk mendiagnosa penyakit diabetes mellitus dengan menerapkan 3 algoritma yaitu Id3, Naïve Bayes, *Smo* dan *Part* dengan akurasi terbaik ialah *SMO* sebesar 77.5%, Naïve Bayes 76.5%, *Decision Tree* J48 75.6%, *PART* 74.4%[11].

Berdasarkan uraian tentang permasalahan diatas, dapat diambil kesimpulan bahwa perlu melakukan pencegahan dan *screening* sejak dini. Sehingga dapat dilakukan tindakan apabila ditemui tanda-tanda gejala penyakit diabetes, dengan cara melakukan *medical checkup* secara rutin [12]. Penelitian ini dapat membantu memberikan keakuratan diagnosis terhadap pasien yang sudah melakukan *checkup* dengan cara menggunakan rekam medis pasien. Penelitian ini dilakukan menggunakan metode data mining klasifikasi yaitu Algoritma Naïve Bayes dan KNN guna mengetahui hasil akhir pemeriksaan yang telah dilakukan. Klasifikasi adalah proses menciptakan sekumpulan model yang berfungsi untuk memaparkan dan memilah kategori pada data ataupun konsepnya [13]. Model berguna untuk memprediksi kategori objek dengan kategori yang tidak diketahui[14]. Penelitian ini juga diharapkan dapat membantu memberikan hasil yang akurat berdasarkan data yang diperoleh sehingga dapat membantu dalam hal pencegahan dan screening penyakit diabetes.

2. METODE

Metode pada penelitian ini menerapkan klasifikasi pada data mining dengan mengimplementasikan Algoritma Naïve Bayes dan KNN dengan mempraktekkan perhitungan dengan tingkat keakurasian, precision, dan recall dengan menggunakan data yang didapatkan dari *provider Kaggle*. Tahapan pada penelitian ini menerapkan proses *Knowlegde Discovery in Databases (KDD)*. *KDD* mempunyai kelebihan yaitu sangat tersusun menggunakan pola yang tepat. Serta di dalam proses *KDD* mengimplementasikan untuk mengeksplor, mengembangkan dan menciptakan model yang sebelumnya tidak pernah dijumpai [15]. *KDD* memilki alur tahapan seperti yang tertera pada Gambar 1 [16].



Gambar 1. Tahapan Alur Penelitian

2.1. Selection

Pada tahap Selection, dilakukan pemilihan data yang sangat relevan dan sesuai dengan tujuan penelitian dari berbagai sumber yang tersedia. Proses pemilihan ini melibatkan kriteria yang ketat untuk memastikan dataset yang digunakan memiliki representasi yang akurat terkait dengan domain penelitian[17]. Misalnya, kita mempertimbangkan variabel-variabel tertentu yang dianggap krusial dan relevan, serta mengidentifikasi data yang tidak relevan atau tidak lengkap untuk dikecualikan dari analisis.

2.2. Preprocessing

Preprocessing merupakan tahapan kritis dalam penelitian ini. Proses ini mencakup langkah-langkah pembersihan data, transformasi, dan pengaturan data agar sesuai untuk analisis lebih lanjut[17]. Dalam pembersihan data, kami mengatasi nilai-nilai yang hilang, outlier, dan inkonsistensi data. Selanjutnya, melakukan transformasi data seperti normalisasi, pengkodean variabel kategori, dan penyatuan data dari berbagai sumber. Preprocessing ini bertujuan untuk memastikan bahwa data yang digunakan dalam proses data mining adalah data yang bersih, konsisten, dan siap untuk diolah.

2.3. Transformation

Tahap Transformasi melibatkan perubahan bentuk atau representasi data untuk meningkatkan relevansi dan kemudahan analisis. Salah satunya menggunakan teknik pengkodean variabel kategori untuk mengubah data kategori menjadi bentuk yang dapat diolah oleh algoritma[17]. Selain itu, juga melakukan penggabungan data dari beberapa atribut untuk menciptakan

https://doi.org/10.25077/TEKNOSI.v9i3.2023.247-254

atribut baru yang lebih informatif. Transformasi ini bertujuan untuk meningkatkan kemampuan algoritma data mining dalam mengidentifikasi pola atau wawasan yang relevan dari data yang telah dipersiapkan.

2.4. Data Mining

Data mining melibatkan pemeriksaan dan analisis data dalam jumlah besar untuk mengidentifikasi pola, tren, informasi tersembunyi, atau wawasan yang berguna [18]. Ia menggunakan teknik komputasi, statistik, dan kecerdasan buatan untuk menghasilkan informasi yang dapat digunakan untuk pengambilan keputusan, peningkatan bisnis, dan pemahaman yang lebih baik tentang data yang ada [19]. Data mining banyak digunakan di berbagai bidang seperti bisnis, sains, kesehatan dan keuangan untuk menemukan informasi berharga yang tidak terlihat pada pandangan pertama. Pada penelitian ini terdapat 2 algoritma yang akan digunakan dalam proses data mining, yaitu:

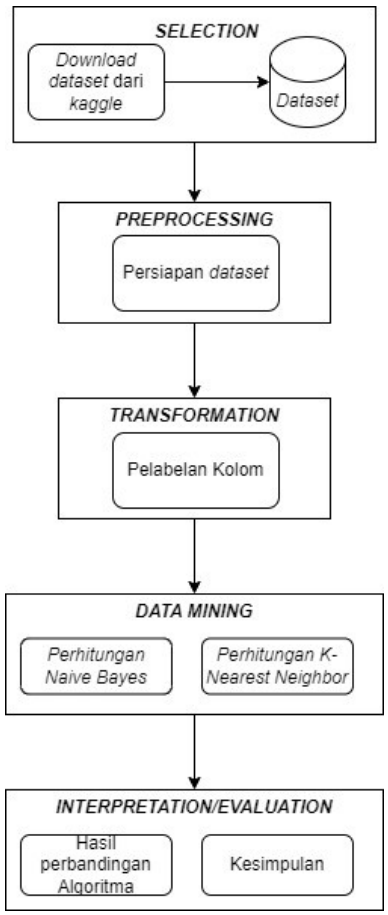
a) Naïve Bayes ialah pengklasifikasi probabilitas yang simple untuk menjumlahkan satu set, peluang dihitung menggunakan penambahan frekuensi dan campuran nilai melalui sekumpulan data yang disediakan[20]. Algoritma Naïve Bayes menerapkan *teorema Bayes* dan memperkirakan bahwa seluruh atribut independen atau independen satu sama lain Variabel kelas[21].
b) K-Nearest Neighbor (KNN)
   Sedangkan *K-Nearest Neighbor (KNN*) ialah algoritma klasifikasi terhadap kumpulan data berlandaskan pengkajian data yang telah terklasifikasi [22]. KNN juga termasuk *supervised learning*, yang mana hasil dari *query instance* yang selesai diklasifikasi berdasarkan kebanyakan jarak terdekat dari kategori yang ada [23]. yang dalam prakteknya mengambil data terdekat "K" (tetangganya) untuk dijadikan acuan dalam proses menentukan kelas dari data yang baru dengan cara mengklasifikasi berdasarkan kedekatan/kemiripannya dengan data yang lain

2.5. Interpretation/Evaluation

Interpretation/Evaluation adalah tahap akhir di mana hasil dari proses data mining dievaluasi dan diinterpretasikan. Ini mencakup analisis temuan, pengidentifikasian wawasan yang bermanfaat, dan penilaian sejauh mana tujuan analisis telah tercapai[17]. Hasil dari tahap ini digunakan untuk pengambilan keputusan, perbaikan proses, atau pemahaman lebih lanjut tentang data yang ditemukan. Interpretation/evaluation merupakan inti dari proses *KDD* untuk menghasilkan nilai tambah dari data.

3. HASIL DAN PEMBAHASAN

Dari tahapan yang sudah dijelaskan maka penelitian ini membawa dalam domain yang kompleks dan mendalam dari data mining, di mana penerapan metode klasifikasi melibatkan Algoritma Naïve Bayes dan KNN. Dalam rangkaian tahapan yang sistematis, mulai dari seleksi data hingga analisis hasil yang mendalam, telah menjalankan sebuah eksplorasi dengan terperinci. Dalam bagian ini akan disampaikan detil dari tahapan pada Gambar 2 dan disampaikan pula hasil temuan dari penelitian ini dengan detil.



Gambar 2. Detail Tahapan

3.1. Selection

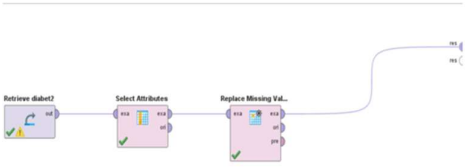
Pada tahap ini adalah proses penyeleksian data yang akan diterapkan terhadap penelitian ini. Data yang diperoleh untuk penelitian ini adalah tentang penyakit diabetes dengan 768 data dengan 9 atribut. Pada Tabel 1 adalah hasil serta penjelasan tiap atribut pada data penyakit diabetes yang didapatkan.

Tabel 1. Penjelasan atribut

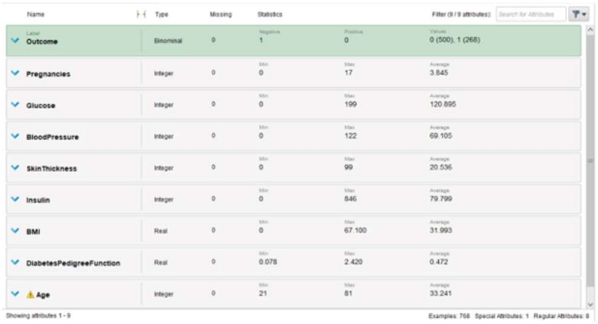
KD	NAMA ATRIBUT	KETERANGAN
P	*Pregnancies*	Hamil berapa kali
G	*Glucose*	Glukosa 2 jam
B	*BloodPressure*	Tensi darah
S	*SkinThickness*	Ketebalan kulit
I	*Insuline*	Insulin 2 Jam
M	*BMI*	Indeks massa tubuh
D	*DiabetesPedigreeFunction*	Fungsi silsilah diabetes
A	*Age*	Usia (dalam hitungan Tahun)
O	*Outcome*	Label

3.2. *Preprocessing*

Pada tahap *preprocessing* ini, data yang telah diseleksi dilakukan cleaning data guna mengetahui *missing value* yang terdapat pada data tersebut. Proses *missing value* dilakukan dengan menggunakan *software RapidMiner*. Pada gambar 2 proses *cleaning* yang dilakukan sedangkan gambar 3 merupakan hasil *cleaning*.



Gambar 3. Proses *Missing Value*



Gambar 4. Hasil *Missing Value*

Setelah melalui proses missing value menggunakan *RapidMiner* pada gambar 4 dapat dikatakan bahwa hasilnya adalah nol, yang berarti bahwa data yang digunakan telah menjadi bersih, tidak ada lagi nilai yang hilang atau kosong dalam *dataset* tersebut, sehingga dapat digunakan untuk proses selanjutnya.

3.3. *Transformation*

Pada tahap ini adalah mentransfromasikan data sehingga memiliki entitas yang jelas. Sehingga data tersebut digunakan untuk melakukan suatu proses data mining dengan mengimplementasikan Algoritma *Naïve Bayes* dan KNN. Namun sebelumnya data harus diubah terlebih dahulu menjadi bilangan biner yaitu 1 dan 0. Bilangan 1 memiliki arti positif diabetes, sedangkan bilangan 0 memiliki arti negatif diabetes. Pada tabel 2 adalah tabel data yang sudah dilakukan transformasi.

Tabel 2. Transformasi Data

NO	P	G	B	S	I	M	D	A	O
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
768	1	93	70	31	0	30.4	0.315	23	0

3.4. *Data Mining*

Proses pengolahan data mining mengimplementasikan Algoritma Naïve Bayes dan KNN dengan bahasa pemrograman *python*. *Python* dapat digunakan untuk mengolah data dengan beberapa metode termasuk Naïve Bayes dan KNN, *library* yang digunakan adalah *numpy, pandas,* . Berikut ini penjelasan proses dari kedua algoritma yang digunakan.

a) Naïve Bayes

Pada proses pengolahan data menggunakan algoritma naïve bayes langkah pertama yaitu mengimport *library* dan *dataset* yang dibutuhkan. Pengolahan data disini menggunakan bahasa pemrograman *python* dan *library* yang dibutuhkan adalah *numpy, pandas, sklearn, matplotlib.* Proses *import library* dan *dataset* dapat dilihat pada gambar 5.

```
# Impor library yang dibutuhkan
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, precision_score, recall_score, roc_auc_score, roc_curve
import matplotlib.pyplot as plt

# Membaca dataset diabetes (gantilah dengan lokasi dan nama file yang sesuai)
diabetes_data = pd.read_csv("dataset_diabetes.csv")
```

Gambar 5. *Import library* dan *dataset naïve bayes*

Setelah mengimport *library* dan *dataset* yaitu memisahkan atribut dalam *dataset*, variabel x digunakan sebagai *input* untuk melatih model dan variabel y sebagai target yang ingin diprediksi oleh model, proses tersebut dapat dilihat pada gambar 6.

```
# Memisahkan fitur (X) dan target (y)
X = diabetes_data.drop("Outcome", axis=1)
y = diabetes_data["Outcome"]
```

Gambar 6. Memisahkan fitur dan target

Sebelum melakukan inisialisasi naïve bayes yaitu memisahkan *dataset* menjadi data *training* 80% dan data *testing* 20% seperti pada gambar 7.

```
# Memisahkan dataset menjadi data training dan data testing (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Skalakan fitur-fitur
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Inisialisasi model Naive Bayes
nb_model = GaussianNB()
```

Gambar 7. Inisialisasi naïve bayes

Setelah melakukan proses inisialisasi yaitu menghitung parameter evaluasi dan menampilkan hasilnya yang mana proses evaluasi diamati melalui parameter yang digunakan yaitu *Accuracy, precision, Recall dan AUC.* Hasil parameter yang digunakan dapat dilihat pada gambar 8.

https://doi.org/10.25077/TEKNOSI.v9i3.2023.247-254

```
# Menghitung parameter evaluasi
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, nb_model.predict_proba(X_test)[:, 1])
```

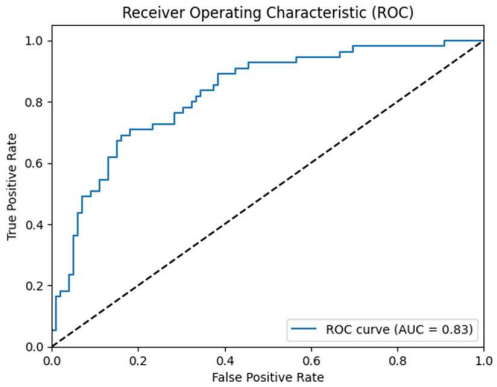
```
# Menampilkan hasil metrik evaluasi
print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
print(f"Recall: {recall:.2f}")
print(f"AUC: {roc_auc:.2f}")
```

```
Accuracy: 0.77
Precision: 0.66
Recall: 0.71
AUC: 0.83
```

Gambar 8. Hasil parameter evaluasi

Selanjutnya, hasil ROC untuk untuk mengukur kualitas dan kinerja model klasifikasi pada berbagai threshold pengklasifikasian yang dapat dilihat pada gambar 9.

```
# Kurva ROC
fpr, tpr, thresholds = roc_curve(y_test, nb_model.predict_proba(X_test)[:, 1])
plt.figure()
plt.plot(fpr, tpr, label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC)')
plt.legend(loc="lower right")
plt.show()
```



Gambar 9. Kurva ROC

b) *K-Nearest Neighbor (KNN)*

Pada proses pengolahan data menggunakan algoritma *k-nearest neighbor (KNN)* langkah yang harus dilakukan pertama yaitu mengimport *library* dan *dataset* yang dibutuhkan. Pengolahan data untuk KNN sama dengan Naïve bayes yaitu menggunakan bahasa pemrograman *python* dan library yang dibutuhkan adalah *numpy, pandas, sklearn, dan matplotlib.* Proses *import library* dan *dataset* dapat dilihat pada gambar 9.

```
# Impor library yang dibutuhkan
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, roc_auc_score, roc_curve
import matplotlib.pyplot as plt
```

```
# Membaca dataset diabetes (gantilah dengan lokasi dan nama file yang sesuai)
diabetes_data = pd.read_csv("dataset_diabetes.csv")
```

Gambar 10. *Import library* dan *dataset KNN*

Setelah melakukan proses *import library* dan *dataset* yaitu melakukan inisialisasi *k-nearest neighbor (KNN),* tetapi sebelum itu harus memisahkan atribut yang akan dijadikan input dan yang akan dijadikan target, atribut yang dipilih yaitu *Outcome* seperti pada gambar 6. Lalu memisahkan *dataset* menjadi data *training 80%* dan data *testing 20%* seperti pada gambar 10. Inilisiasi model KNN menggunakan KNN K=7

```
# Memisahkan dataset menjadi data training dan data testing (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Skalakan fitur-fitur
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
# Inisialisasi model K-NN dengan k=n
knn_model = KNeighborsClassifier(n_neighbors=3)
```

Gambar 11. Inisialisasi k-nearest neighbor (KNN)

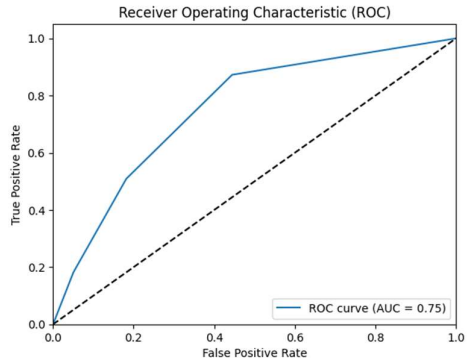
Setelah melakukan proses inisialisasi adalah menghitung parameter evaluasi serta menampilkan hasilnya. Proses evaluasi yang diamati sama dengan algoritma naïve bayes yang dapat dilihat pada gambar 8. Untuk mengetahui hasil evaluasi KNN=3 dan kurva ROC guna mengukur kualitas dan kinerja model klasifikasi pada berbagai threshold pengklasifikasian dapat dilihat pada gambar 12.

```
# Menampilkan hasil parameter evaluasi KNN=3
print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
print(f"Recall: {recall:.2f}")
print(f"AUC: {roc_auc:.2f}")
```

```
Accuracy: 0.71
Precision: 0.61
Recall: 0.51
AUC: 0.75
```

Gambar 12. Hasil parameter KNN=3

```
# Kurva ROC KNN=3
fpr, tpr, thresholds = roc_curve(y_test, knn_model.predict_proba(X_test)[:, 1])
plt.figure()
plt.plot(fpr, tpr, label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC)')
plt.legend(loc="lower right")
plt.show()
```



Gambar 13. Kurva ROC KNN=3

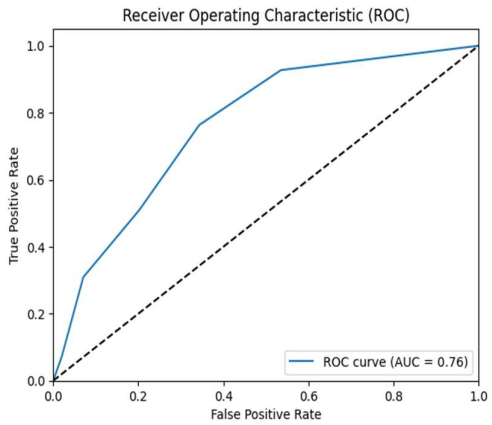
Setelah melakukan perhitungan KNN=3 maka dilanjutkan dengan KNN=5 dengan parameter yang sama.

```
# Menampilkan hasil parameter evaluasi KNN=5
print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
print(f"Recall: {recall:.2f}")
print(f"AUC: {roc_auc:.2f}")
```

```
Accuracy: 0.69
Precision: 0.58
Recall: 0.51
AUC: 0.76
```

Gambar 14. Hasil parameter KNN=5

```
# Kurva ROC KNN=3
fpr, tpr, thresholds = roc_curve(y_test, knn_model.predict_proba(X_test)[:, 1])
plt.figure()
plt.plot(fpr, tpr, label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC)')
plt.legend(loc="lower right")
plt.show()
```



Gambar 15. Kurva ROC KNN=5

Selanjutnya, hasil menghitungan parameter evaluasi untuk dan ROC KNN=7 yang dapat dilihat pada Gambar 16 dan 17.

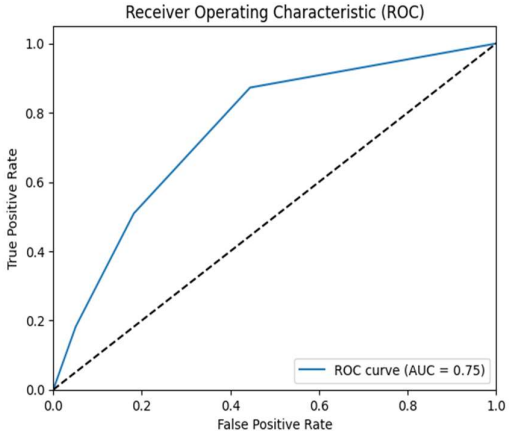
```
# Menghitung parameter evaluasi
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, knn_model.predict_proba(X_test)[:, 1])
```

```
# Menampilkan hasil parameter evaluasi
print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
print(f"Recall: {recall:.2f}")
print(f"AUC: {roc_auc:.2f}")
```

```
Accuracy: 0.71
Precision: 0.61
Recall: 0.51
AUC: 0.75
```

Gambar 16. Hasil parameter evaluasi KNN=7

```
# Kurva ROC
fpr, tpr, thresholds = roc_curve(y_test, knn_model.predict_proba(X_test)[:, 1])
plt.figure()
plt.plot(fpr, tpr, label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC)')
plt.legend(loc="lower right")
plt.show()
```



Gambar 17. Kurva ROC KNN=7

3.5. Interpretation/Evaluation

Setelah melakukan proses pengolahan data menggunakan *python* didapatkan hasil untuk setiap metode. Pada proses diatas terdapat menggunakan split data, yaitu data training sebesar 80 % dan data testing sebesar 20%. Hasil akhir proses perhitungan diatas dengan menggunakan *confusion matrix* antara lain *Accuracy, Precision, Recall* yang tertera di tabel 3.

Hasil pada tabel 3 menunjukkan bahwa metode Naïve Bayes memiliki hasil akurasi lebih unggul daripada KNN, namun kedua algoritma ini memiliki keterbatasan yaitu naïve bayes menunjukkan keterbatasan mengidentifikasi pola yang rumit dalam dataset medis yang beragam sedangkan KNN memiliki keterbatasan pada pemilihan parameter *k* seperti parameter yang teralu rendah dan tinggi mempengaruhi kinerja model.

Sedangkan prospek dalam aplikasi nyatanya dapat digunakan sebagai alat keputusan klinis meski tingkat akurasi KNN lebih rendah dibandingkan dengan Naïve Bayes, namun keduanya masih dapat diintegrasikan ke dalam alat pendukung keputusan

https://doi.org/10.25077/TEKNOSI.v9i3.2023.247-254

klinis. Semakin tinggi akurasi Naïve Bayes, semakin akurat hasil yang dapat diberikan untuk memastikan diagnosis, sedangkan KNN dapat memberikan informasi tambahan, serta pengembangan fitur tambahan yang lebih spesifik terhadap medis atau dataset diabetes. Namun penting untuk dicatat bahwa hasil metode ini hanyalah prediksi dan bukan diagnosis akhir, kedua metode ini dapat memberikan kontribusi terhadap gambaran rinci risiko diabetes pada tingkat individu, dan keterbatasannya perlu dipahami dalam konteks penerapan di dunia nyata.

Tabel 3. Hasil Klasifikasi

Klasifikasi	Accuracy	Precision	Recall	AUC
Naïve Bayes	77%	66%	71%	0.83
KNN=3	71%	61%	51%	0.75
KNN=5	69%	58%	51%	0.76
KNN=7	68%	56%	49%	0.75

Penelitian ini memiliki perbedaan dengan penelitian terdahulu yang sudah dijelaskan yaitu berfokus pada mengkomparisakan kinerja 2 algoritma yang berbeda dalam pengolahan data, selain itu penelitian ini mengintegrasikan dua *platform* analisis yakni *RapidMiner* dan *Python* untuk mendukung pemrosesan data yang lebih komprhensif. Pada implementasi metode KNN yang ditingkatkan dengan menampilkan informasi 3 paramater *k* terdekat untuk menentukan *k* mana yang memiliki kontribusi signifikan pada analisis data yang dilakukan.

4. KESIMPULAN

Berdasarkan hasil akhir keseluruhan dalam melakukan perbandingan terhadap 2 algorima yaitu Naïve Bayes dan *K-nearest Neighbor (KNN)* yaitu metode Naïve Bayes memiliki Accuracy sebesar 77%, Precision 66% dan untuk recall adalah sebesar 71%. Sedangkan metode K-Nearest Neighbor (KNN) dengan K=3, K=5, dan K=7 adalah K=3 lebih unggul dari segi *accuracy, precision* dan *recal*, sedangkan untuk nilai AUC K=5 lebih unggul yaitu sebesar 0.76 dibandingkan K=3 dan K7. Keunggulan ini diukur berdasarkan parameter yang disebutkan pada tabel 3, tabel tersebut menunjukkan bahwa algoritma naïve bayes lebih unggul dibandingkan KNN sehingga naïve bayes menjadi pilihan yang lebih layak karena menghasilkan akurasi yang lebih unggul dan dapat digunakan sebagai alat bantu dalam prediksi dan identifikasi dini pada penyakit diabetes sehingga memudahkan tenaga medis dalam memberikan diagnosis dan perawatan yang lebih tepat waktu.

DAFTAR PUSTAKA

[1] N. Nurdiana and A. Algifari, "Studi Komparasi Algoritma Id3 Dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *INFOTECH J.*, vol. 6, no. 2, pp. 18–23, 2020, [Online]. Available: https://ejournal.unma.ac.id/index.php/infotech/article/view/816

[2] Y. Sinatrya and L. A. Wulandhari, "Deteksi Diabetes Melitus Untuk Wanita Dan Penyusunan Menu Sehat Dengan Pendekatan Adaptive Neuro Fuzzy Inference System (Anfis) Dan Algoritma Genetika (Ga)," *J. Tek. Inform.*, vol. 12, no. 1, pp. 39–58, 2019, doi: 10.15408/jti.v12i1.9578.

[3] F. Fitriyani, "Prediksi Diabetes Menggunakan Algoritma Naive Bayes dan Greedy Forward Selection," *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 2, pp. 61–69, 2021, doi: 10.25077/teknosi.v7i2.2021.61-69.

[4] D. A. NAWANGNUGRAENI, "Sistem Pakar Berbasis Android untuk Diagnosis Diabetes Melitus dengan Metode Forward Chaining," *Komputika J. Sist. Komput.*, vol. 10, no. 1, pp. 19–27, 2021, doi: 10.34010/komputika.v10i1.3553.

[5] M. Yusa, E. Utami, and E. T. Luthfi, "Analisis Komparatif Evaluasi Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes," *J. Buana Inform.*, vol. 7, no. 4, pp. 293–302, 2016, doi: 10.24002/jbi.v7i4.770.

[6] S. Rokhanah, A. Hermawan, and D. Avianto, "Pengaruh Principal Component Analysis Pada Naïve Bayes dan K-Nearest Neighbor Untuk Prediksi Dini Diabetes Melitus Menggunakan Rapidminer," *EVOLUSI J. Sains dan Manaj.*, vol. 11, no. 1, 2023, doi: 10.31294/evolusi.v11i1.14728.

[7] R. Oktaria, M. Komarudin, and M. A. Muda, "Analisa Klasifikasi Kualitas Mahasiswa Lulusan Berdasarkan Jalur Penerimaan Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Lampung)," *J. Tek. Inform.*, vol. 12, no. 2, pp. 183–192, 2019, doi: 10.15408/jti.v12i2.11171.

[8] B. D. Meilani, N. Susanti, J. T. Informatika, F. T. Informasi, I. Teknologi, and A. Tama, "Akurasi Data Mining Untuk Menghasilkan Pola Kelulusan Mahasiswa dengan Metode NAÏVE BAYES," *J. Sist. Inf. Univ. Suryadarma*, vol. 3, no. 2, pp. 182–189, 2014, doi: 10.35968/jsi.v3i2.66.

[9] J. I. Marzuki, K. Mataram, and N. T. Bar, "Komparasi Akurasi Metode Correlated Naive Bayes Classifier Dan Naive Bayes Classifier Untuk Diagnosis Penyakit Diabetes Hairani , Gibran Satya Nugraha , Mokhammad Nurkholis Abdillah , Muhammad Innuddin InfoTekJar (Jurnal Nasional Informatika dan Teknolog," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 3, no. 1, pp. 6–11, 2018.

[10] F. Yunita, "Sistem Klasifikasi Penyakit Diabetes Mellitus Menggunakan Metode K-Nearest Neighbor (K-NN)," *Bappeda*, vol. 2, pp. 223–230, 2016.

[11] Nurahman and Prihandoko, "Perbandingan Hasil Analisis Teknik Data Mining 'Metode Decision Tree, Naive Bayes, Smo Dan Part' Untuk Mendiagnosa Penyakit Diabetes Mellitus," *J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. Vol. 4 No., 2019, [Online]. Available: https://ejournal.unitomo.ac.id/index.php/inform/article/view/1403/pdf

[12] D. Setyawan and A. Suradi, "Implementasi Web Service Dan Analisis Kinerja Algoritma Klasifikasi Data Mining Untuk Memprediksi Diabetes Mellitus," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 8, no. 2, p. 701, 2017, doi: 10.24176/simet.v8i2.1584.

[13] Y. N. Dewi and F. A. Sariasih, "Metode Sample Bootstrapping Untuk Meningkatkan Performa Algoritma Naive Bayes Pada Citra Tunggal Pap Smear," *J. Tek. Inform.*, vol. 12, no. 1, pp. 1–10, 2019, doi: 10.15408/jti.v12i1.11031.

[14] L. N. Rani, "Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit," *INOVTEK Polbeng - Seri Inform.*, vol. 1, no. 2, p. 126, 2016, doi: 10.35314/isi.v1i2.131.

[15] E. Luthfi and A. W. Wijayanto, "Analisis perbandingan metode hirearchical , k-means , dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia Indonesia

Comparative analysis of hirearchical , k-means , and k-medoids clustering and methods in grouping Indonesia ' s human," *Inovasi*, vol. 17, no. 4, pp. 770–782, 2021.

[16] S. Handoko, F. Fauziah, and E. T. E. Handayani, "Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering," *J. Ilm. Teknol. dan Rekayasa*, vol. 25, no. 1, pp. 76–88, 2020, doi: 10.35760/tr.2020.v25i1.2677.

[17] A. Franseda, W. Kurniawan, S. Anggraeni, and W. Gata, "Integrasi Metode Decision Tree dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 3, p. 282, 2020, doi: 10.26418/justin.v8i3.40982.

[18] P. N. Harahap and S. Sulindawaty, "Implementasi Data Mining Dalam Memprediksi Transaksi Penjualan Menggunakan Algoritma Apriori (Studi Kasus PT.Arma Anugerah Abadi Cabang Sei Rampah)," *Matics*, vol. 11, no. 2, p. 46, 2020, doi: 10.18860/mat.v11i2.7821.

[19] Mardalius, "Pengelompokan Data Penjualan Aksesoris Menggunakan Algoritma K-Means," *Jurteksi*, vol. IV, no. 2, pp. 401–411, 2018.

[20] N. Buslim, L. K. Oh, M. H. Athallah Hardy, and Y. Wijaya, "Comparative Analysis of KNN, Naïve Bayes and SVM Algorithms for Movie Genres Classification Based on Synopsis.," *J. Tek. Inform.*, vol. 15, no. 2, pp. 169–177, 2022, doi: 10.15408/jti.v15i2.29302.

[21] E. Fitriani, "Perbandingan Algoritma C4.5 Dan Naïve Bayes Untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan," *Sistemasi*, vol. 9, no. 1, p. 103, 2020, doi: 10.32520/stmsi.v9i1.596.

[22] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, "Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi," *Sistemasi*, vol. 10, no. 2, p. 445, 2021, doi: 10.32520/stmsi.v10i2.1303.

[23] I. Moment, "Klasifikasi Daun Padi dengan K-Nearest Neighbor Berdasarkan Fitur Warna dan Invariant Moment Rice Classification with K-Nearest Neighbor based on Color Feature," vol. 12, no. September, pp. 660–674, 2023.

BIODATA PENULIS

Qonitah Alia Puteri

Qonitah Alia Puteri lahir di Sumenep 30 Desember 2000, merupakan seorang mahasiswa Program Studi S1 Sistem Informasi Universitas Dinamika, Surabaya.

Tri Sagirani, S.Kom., M.MT.

Tri Sagirani adalah seorang peneliti dan dosen di Program Studi Sistem Informasi, Universitas Dinamika, Surabaya, Indonesia. Telah menempuh pendidikan Strata satu pada program studi Manajemen Informatika di STIKOM Surabaya dan memperoleh gelar Magister Manajemen Teknologi di Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Saat ini, memiliki fokus riset pada bidang interaksi manusia dan komputer, bidang *User Interface/ User Exp*erience (UI/UX). Publikasi yang dihasilkan terkait dengan pemanfaatan teknologi informasi dibeberapa bidang, baik bidang kesehatan, teknologi pendidikan, baik pendidikan untuk jenjang reguler maupun pendidikan khusus.

Julianto Lemantara, S.Kom., M.Eng.

Julianto Lemantara, adalah seorang peneliti dan dosen di Program Studi S1 Sistem Informasi, Universitas Dinamika. Telah lulus dari STMIK Surabaya pada tahun 2009 dengan gelar Sarjana Komputer di bidang sistem Informasi dan melanjutkan studi dan menyelesaikan gelar Master di bidang Teknologi Informasi dari Universitas Gadjah Mada pada tahun 2013. Saat ini, memiliki minat penelitian di bidang sistem informasi, rekayasa perangkat lunak, sistem pendukung keputusan, dan penambangan data. Hingga tahun 2023, beliau telah mempublikasikan karya penelitian di Jurnal dan Prosiding Internasional terindeks Scopus.