CLASSIFICATION

ANGGOTA KELOMPOK

FARAH NASYWA (2208107010051)

IWANI KHAIRINA (2208107010078)

DINDA MAHARANI (2208107010081)

Kemajuan teknologi dalam bidang kecerdasan buatan, khususnya machine learning, telah membuka peluang besar dalam dunia medis, termasuk untuk prediksi risiko kesehatan. Salah satu penyakit yang memerlukan penanganan dini adalah gagal jantung, karena berisiko tinggi menyebabkan kematian.

Dataset Heart Failure Prediction menyediakan data klinis pasien yang dapat dianalisis menggunakan metode klasifikasi seperti KNN, Naive Bayes, Logistic Regression, dan Decision Tree. Dengan menerapkan algoritma tersebut, kita dapat membangun model prediktif untuk memperkirakan kemungkinan kematian pasien.

Studi ini bertujuan mengevaluasi dan membandingkan performa model-model klasifikasi dalam memprediksi kematian pasien gagal jantung, sehingga dapat digunakan sebagai dasar pengambilan keputusan medis yang lebih tepat.

DATASET

Dataset yang digunakan adalah Heart Failure Prediction, yang berisi data rekam medis dari 5000 pasien yang mengalami gagal jantung. Data ini dikumpulkan selama periode tindak lanjut, dengan tujuan utama untuk memprediksi apakah pasien akan meninggal selama masa tersebut. Dataset ini memiliki 13 fitur klinis seperti usia, tekanan darah, kadar kreatinin, dan lainnya, yang relevan dalam menilai kondisi kesehatan pasien.

Heart Failure Clinical Records Dataset

13 FITUR KLINIS

1) Age (usia): Usia pasien dalam tahun.

2) Anaemia (anemia): Menunjukkan apakah pasien mengalami penurunan sel darah merah atau hemoglobin, yang mempengaruhi pengangkutan oksigen.

3) Creatinine Phosphokinase (Kreatin kinase): Level enzim CPK dalam darah (mcg/L), yang menunjukkan kerusakan otot, khususnya otot jantung.

4) Diabetes: Menunjukkan apakah pasien menderita diabetes.

5) Ejection Fraction (Fraksi ejeksi): Persentase darah yang keluar dari jantung setiap kontraksi, dan penting untuk menilai fungsi jantung (dalam persen).

6) High Blood Pressure (Hipertensi): Menunjukkan apakah pasien memiliki tekanan darah tinggi.

7) Platelets (Trombosit): Jumlah trombosit dalam darah (kilotrombosit/mL), yang menunjukkan kemampuan pembekuan darah.

8) Sex (Jenis kelamin): Menunjukkan apakah pasien pria atau wanita.

9) Serum Creatinine (Kreatinin serum): Level kreatinin dalam darah (mg/dL), digunakan untuk menilai fungsi ginjal.

10) Sodium Serum (Serum Sodium): Level sodium dalam darah (mEq/L), yang penting untuk menjaga keseimbangan cairan.

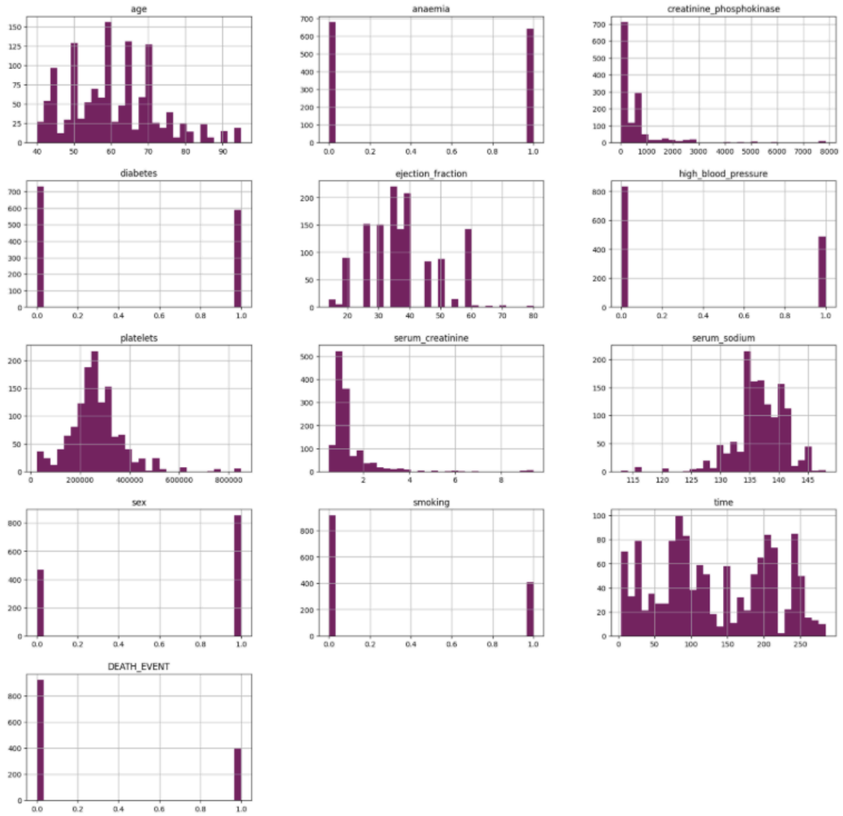
11) Smoking (Merokok): Menunjukkan apakah pasien merokok atau tidak. 12) Time (Waktu): Periode tindak lanjut dalam hari.

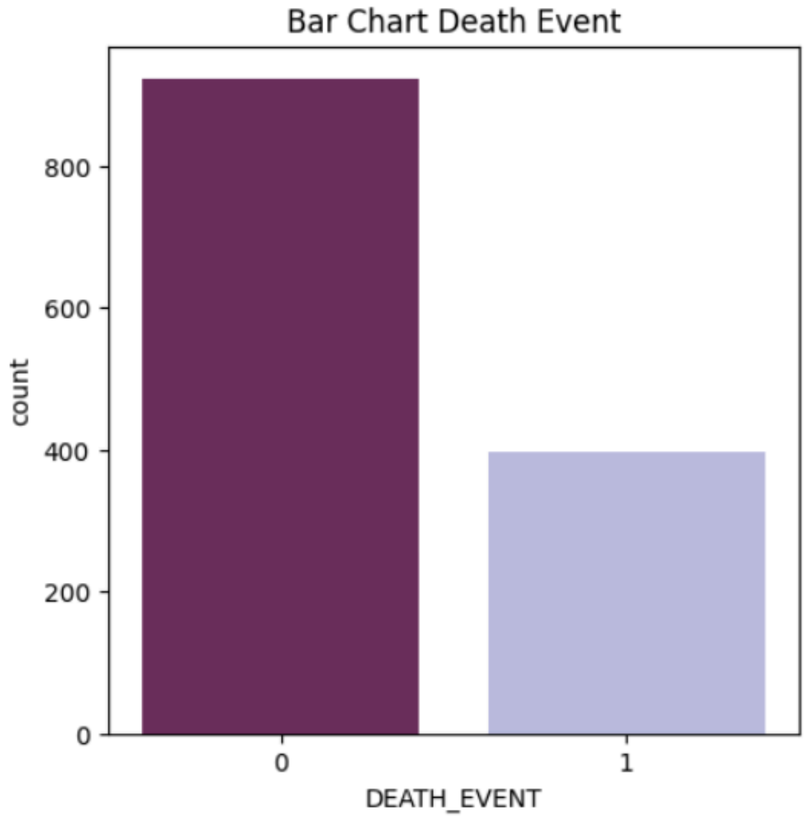
13) DEATH_EVENT (Kematian): Menunjukkan apakah pasien meninggal selama masa tindak lanjut (1 untuk meninggal, 0 untuk masih hidup).

PREPROSESSING DATA

- Menghapus fitur yang kurang relevan atau redundan: anaemia, diabetes, high_blood_pressure, sex, smoking.
- Mengecek apakah terdapat data kosong (missing values).
- Menghapus data duplikat jika ada.
- Melihat ringkasan statistik dari dataset dengan .describe().
- Menampilkan informasi struktur dataset dengan .info().
- Membuat histogram dari semua fitur numerik untuk memahami distribusi datanya.

Visualisasi Distribusi Data





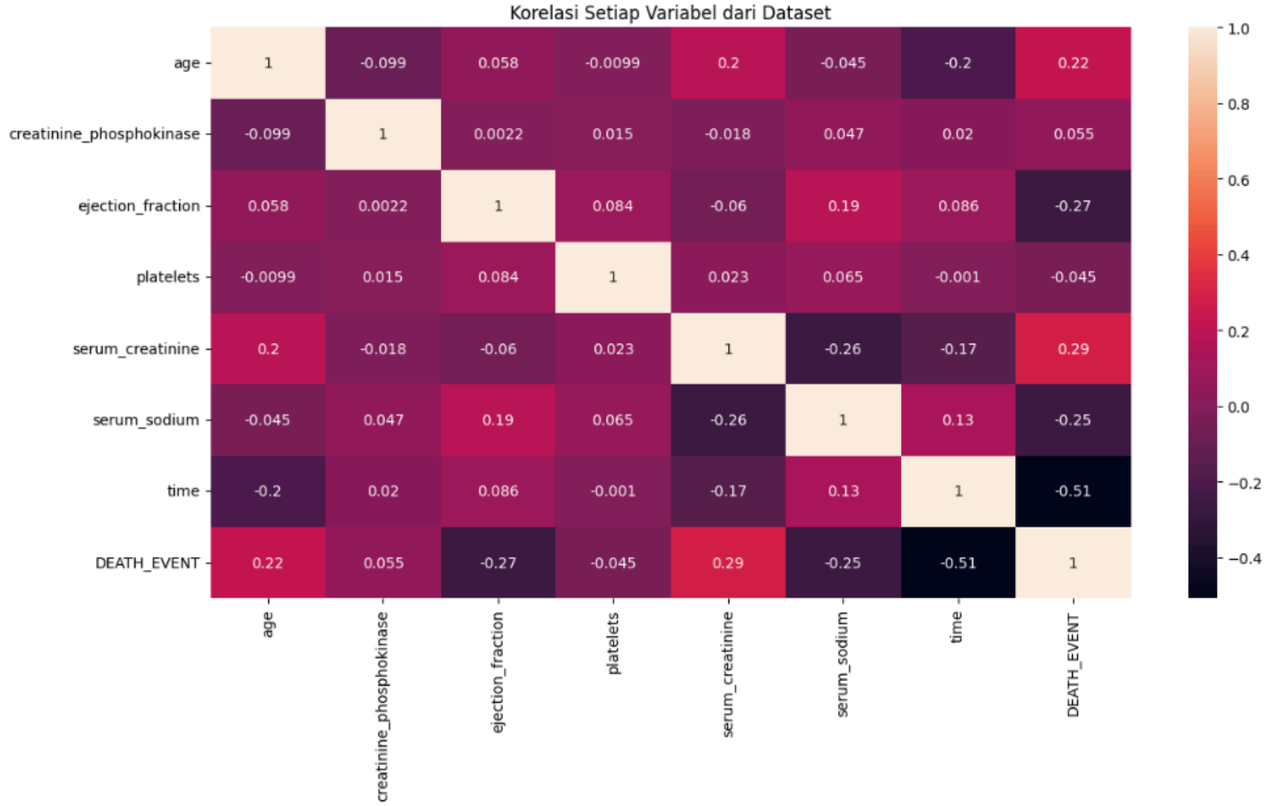
Korelasi Antar Variabel

```
Informasi Dataset:
<class 'pandas.core.frame.DataFrame'>
Index: 1320 entries, 0 to 4972
Data columns (total 8 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       1320 non-null   float64
 1   creatinine_phosphokinase  1320 non-null   int64
 2   ejection_fraction         1320 non-null   int64
 3   platelets                 1320 non-null   float64
 4   serum_creatinine          1320 non-null   float64
 5   serum_sodium              1320 non-null   int64
 6   time                      1320 non-null   int64
 7   DEATH_EVENT               1320 non-null   int64
dtypes: float64(3), int64(5)
memory usage: 92.8 KB
None

Statistik Deskriptif:
                age  creatinine_phosphokinase  ejection_fraction  \
count  1320.000000                1320.000000        1320.000000
mean     60.587377                 576.135606          37.881818
std      11.913538                 970.630878          11.572547
min      40.000000                  23.000000          14.000000
25%      50.000000                 115.000000          30.000000
50%      60.000000                 249.000000          38.000000
75%      69.000000                 582.000000          45.000000
max      95.000000                7861.000000          80.000000

           platelets  serum_creatinine  serum_sodium         time  DEATH_EVENT
count    1320.000000       1320.000000   1320.000000  1320.000000  1320.000000
mean   263751.982189          1.356447    136.665909   132.678788     0.300758
std    106345.010143          0.998924      4.380990    77.779493     0.458761
min     25100.000000          0.500000    113.000000     4.000000     0.000000
25%    208000.000000          0.900000    134.000000    74.000000     0.000000
50%    263358.030000          1.100000    137.000000   119.500000     0.000000
75%    310000.000000          1.300000    140.000000   206.000000     1.000000
max    850000.000000          9.400000    148.000000   285.000000     1.000000

Missing Values:
age                         0
creatinine_phosphokinase    0
ejection_fraction           0
platelets                   0
serum_creatinine            0
serum_sodium                0
time                        0
DEATH_EVENT                 0
dtype: int64
```



Training dan Testing

Langkah ini bertujuan untuk memisahkan data menjadi training set dan testing set:

- X berisi semua fitur kecuali kolom terakhir (target), dan y berisi kolom target DEATH_EVENT.
- Data dibagi dengan rasio 90% untuk pelatihan (train) dan 10% untuk pengujian (test) menggunakan train_test_split() dari sklearn.
- Parameter shuffle=True memastikan data diacak sebelum pembagian.
- stratify=y digunakan agar proporsi kelas pada target tetap seimbang di antara data train dan test.
- random_state=42 memastikan hasil pembagian data bisa direproduksi (reproducible).

- `X_train` dan `y_train` : data latih
- `X_test` dan `y_test` : data uji

```
X = data.iloc[:, :-1].values
y = data.iloc[:, 7].values
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.1,shuffle=True, stratify=y, random_state=42)
```

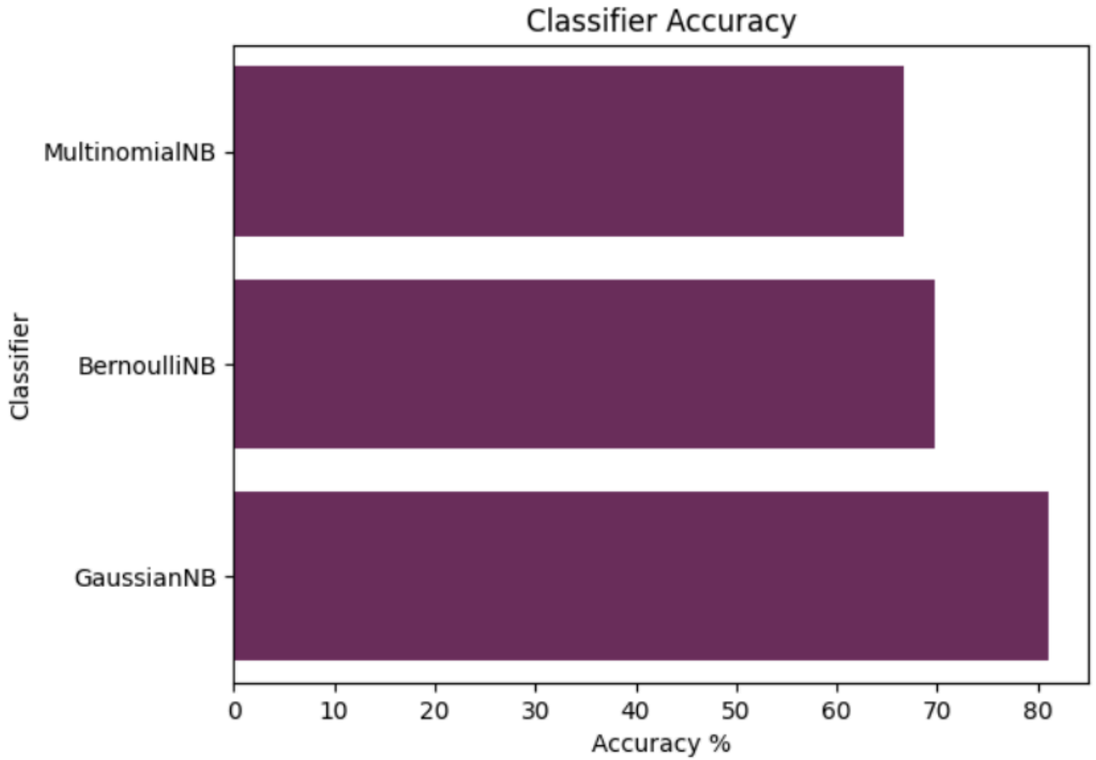
```
print("Jumlah baris X_train:", len(X_train))
print("Jumlah baris X_test:", len(X_test))
print("Jumlah baris y_train:", len(y_train))
print("Jumlah baris y_test:", len(y_test))
```

```
Jumlah baris X_train: 1188
Jumlah baris X_test: 132
Jumlah baris y_train: 1188
Jumlah baris y_test: 132
```

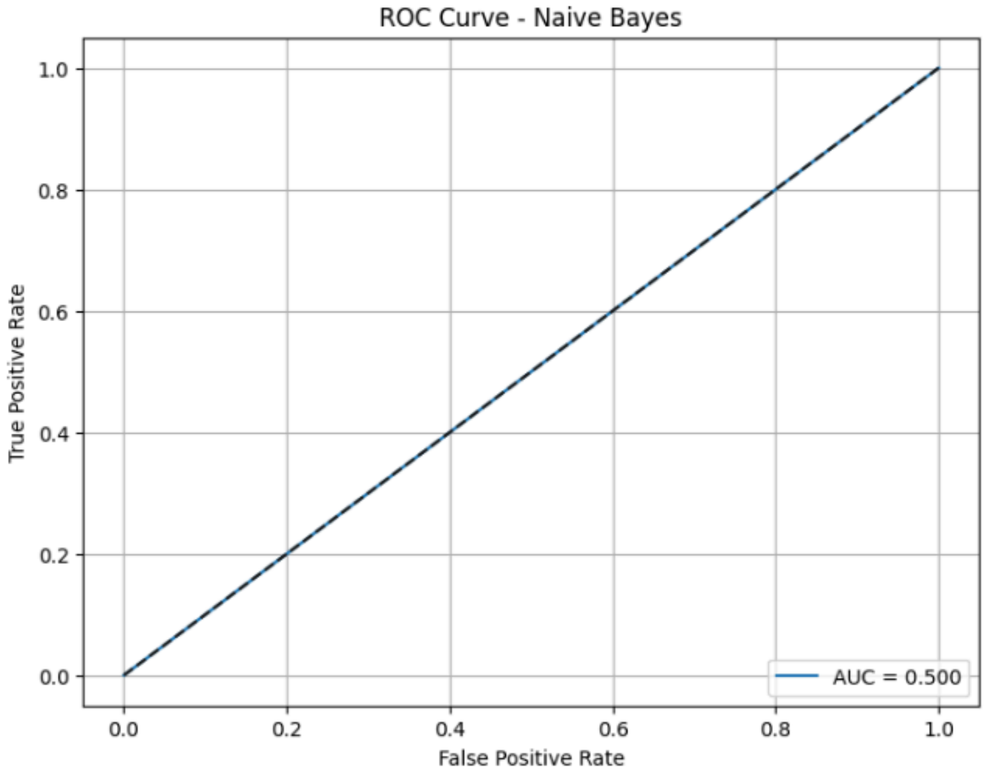
1. Naive Bayes Classifier

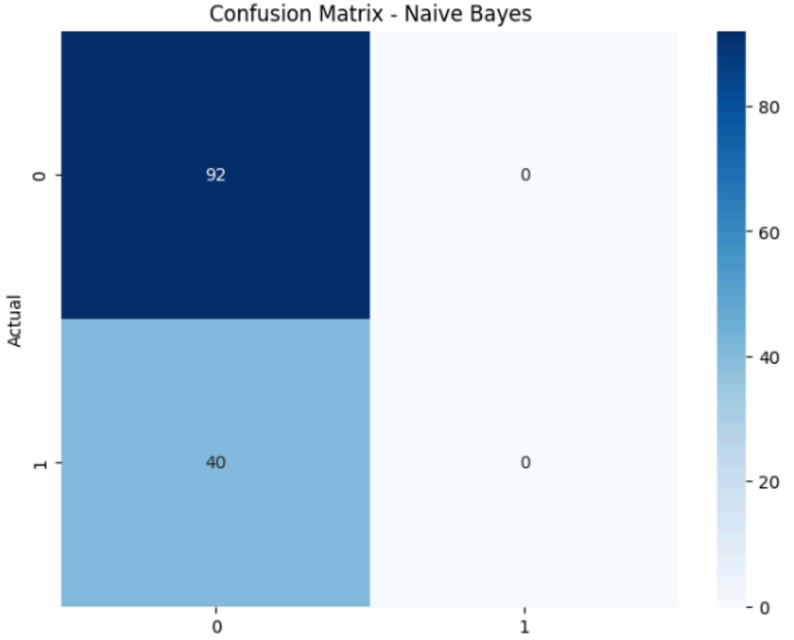
Pada bagian ini, dilakukan evaluasi dan perbandingan kinerja dari tiga varian algoritma Naive Bayes:

```
==============================
MultinomialNB
*Results*
Accuracy: 66.6667%
==============================
==============================
BernoulliNB
*Results*
Accuracy: 69.6970%
==============================
==============================
GaussianNB
*Results*
Accuracy: 81.0606%
==============================
```



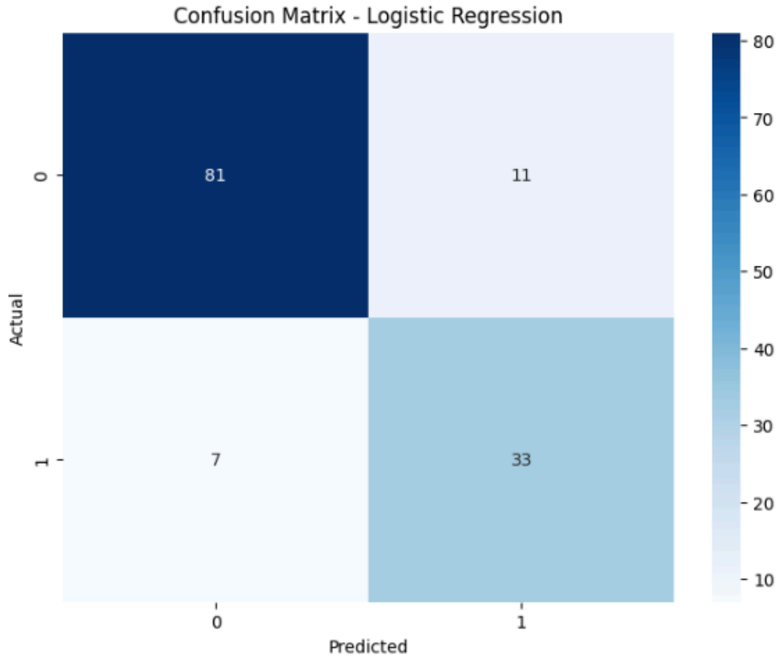
1. Naive Bayes Classifier



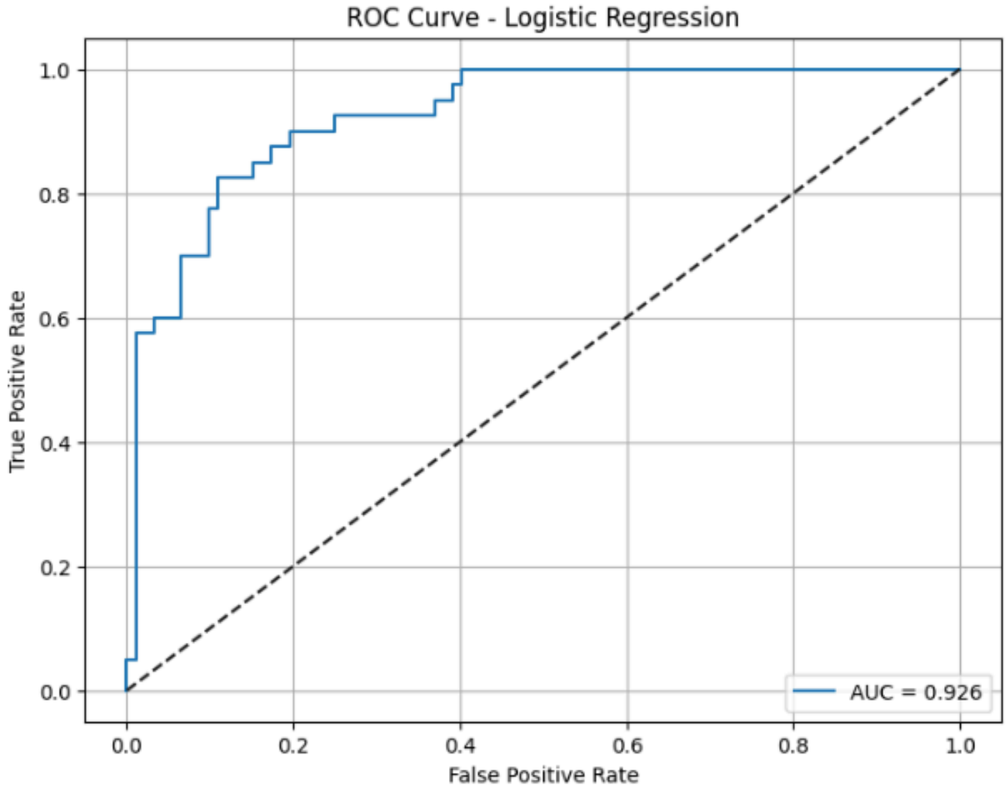


	precision	recall	f1-score	support
0	0.70	1.00	0.82	92
1	0.00	0.00	0.00	40
accuracy			0.70	132
macro avg	0.35	0.50	0.41	132
weighted avg	0.49	0.70	0.57	132

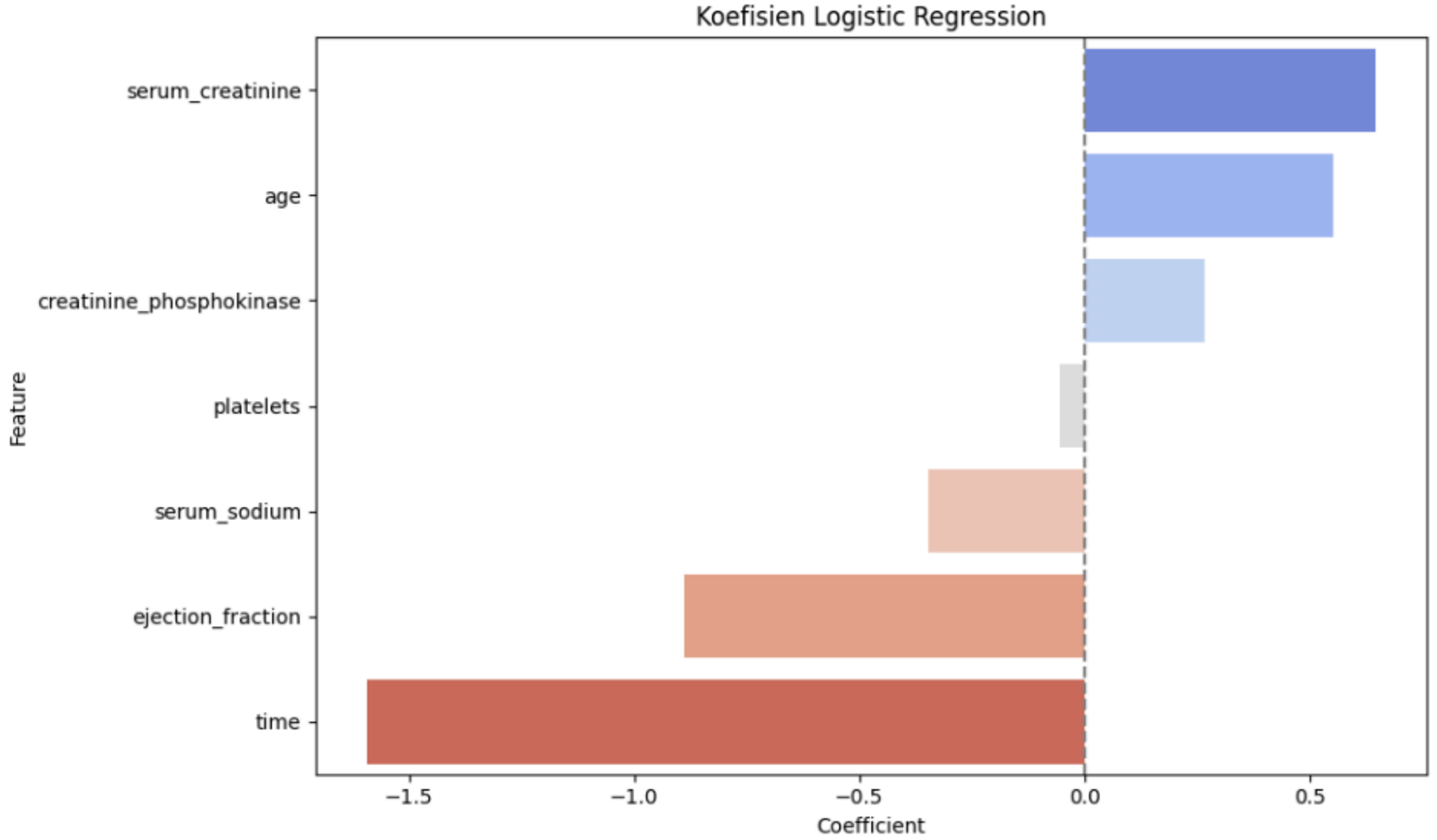
2. Logistic Regression



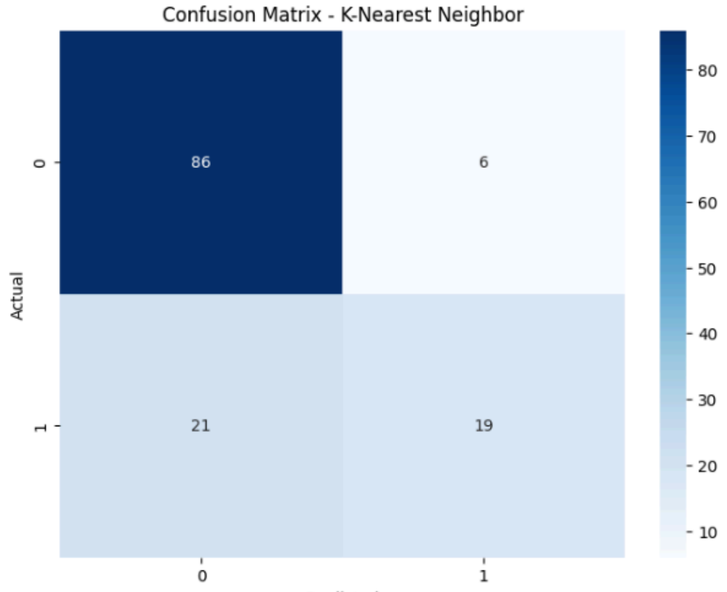
	precision	recall	f1-score	support
0	0.92	0.88	0.90	92
1	0.75	0.82	0.79	40
accuracy			0.86	132
macro avg	0.84	0.85	0.84	132
weighted avg	0.87	0.86	0.87	132



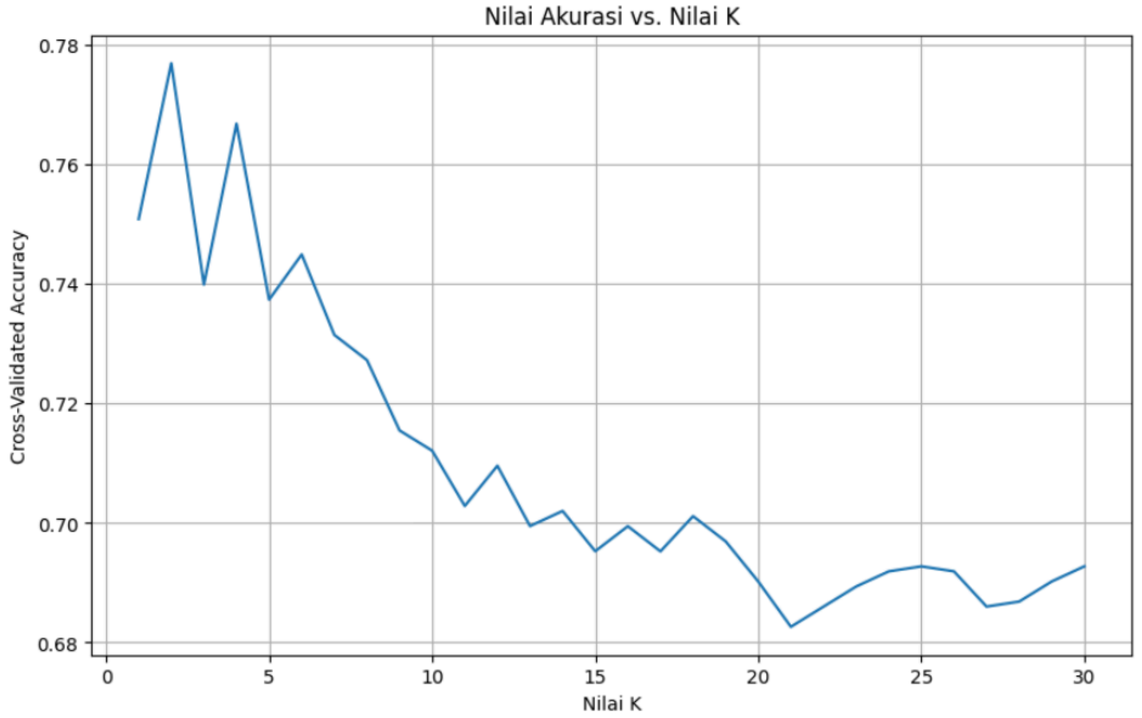
2. Logistic Regression



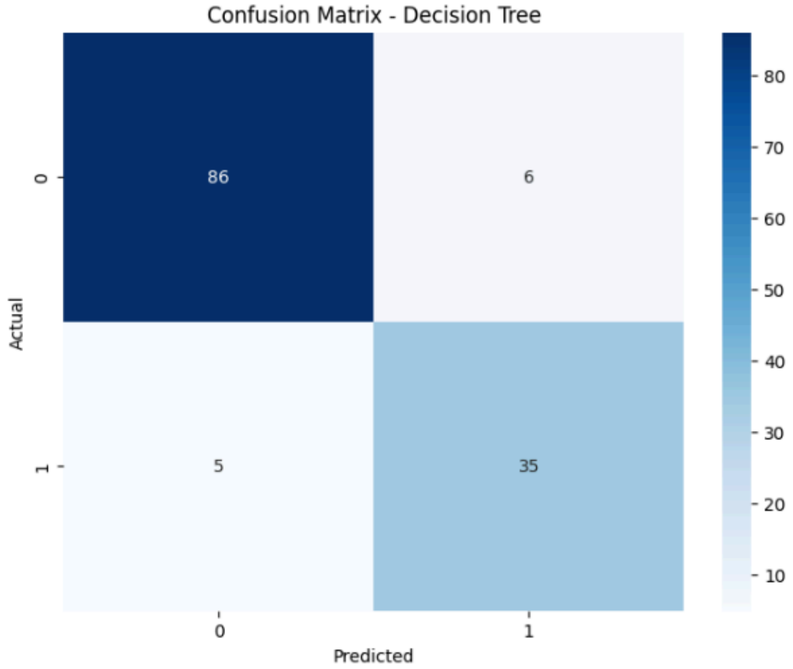
3. K-Nearest Neighbors (KNN)



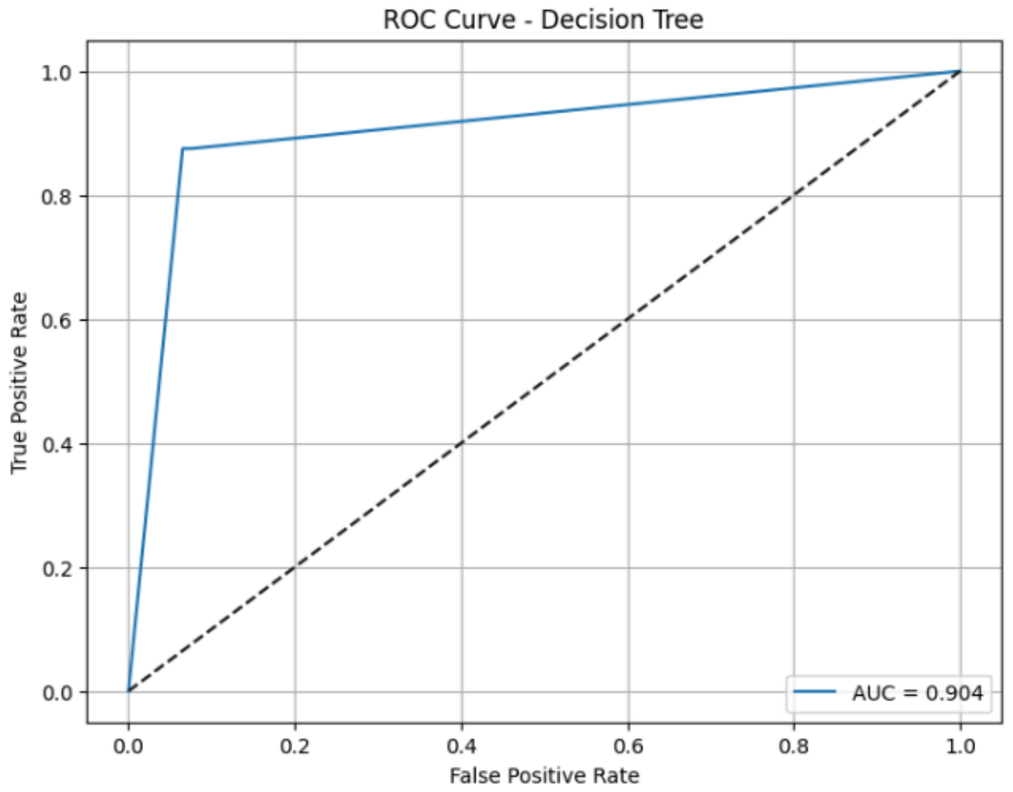
	precision	recall	f1-score	support
0	0.80	0.93	0.86	92
1	0.76	0.47	0.58	40
accuracy			0.80	132
macro avg	0.78	0.70	0.72	132
weighted avg	0.79	0.80	0.78	132



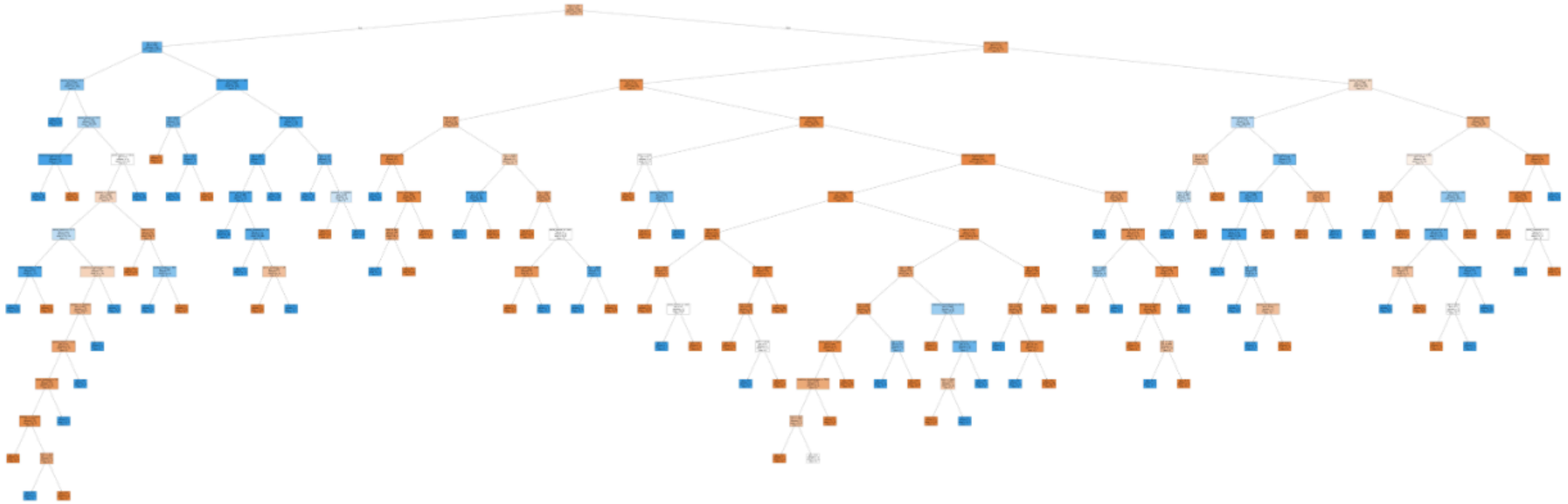
4. Decision Tree



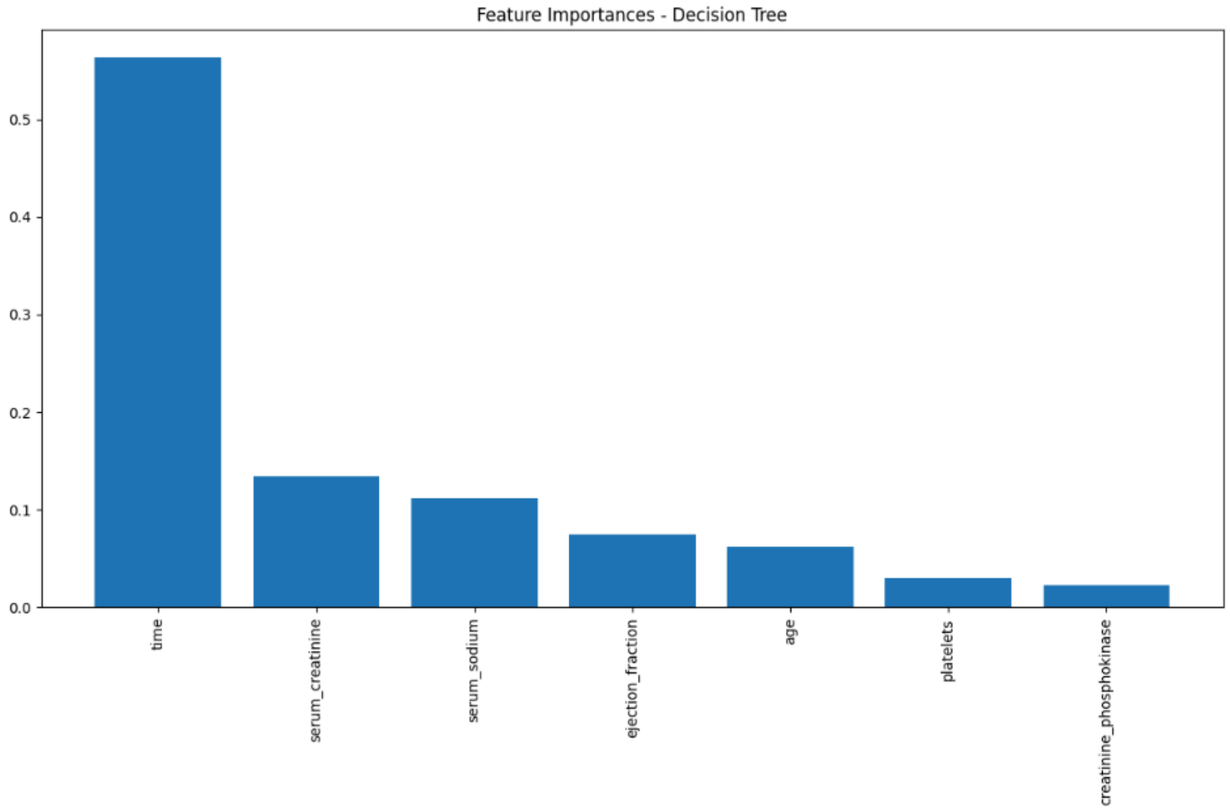
	precision	recall	f1-score	support
0	0.95	0.93	0.94	92
1	0.85	0.88	0.86	40
accuracy			0.92	132
macro avg	0.90	0.90	0.90	132
weighted avg	0.92	0.92	0.92	132



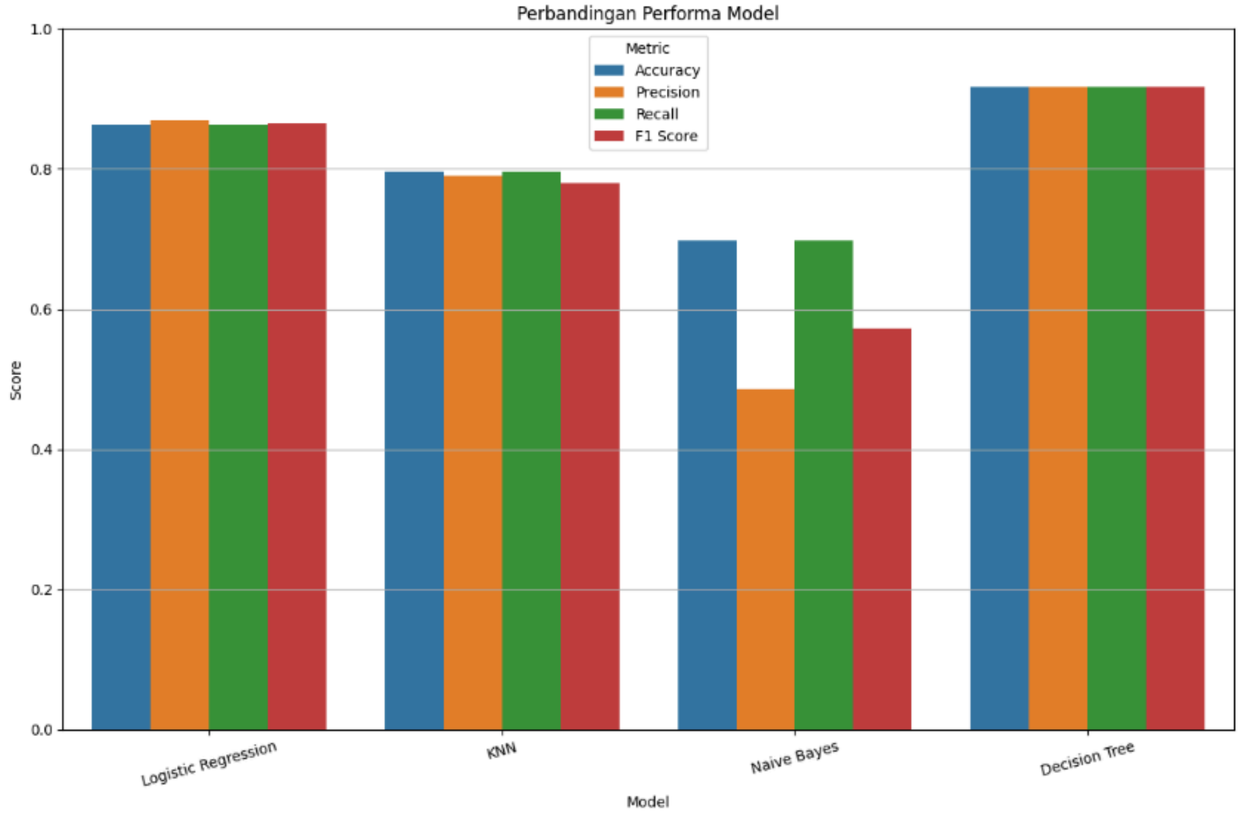
4. Decision Tree



4. Decision Tree



Visualisasi Perbandingan Performa Model



Kesimpulan

- Model terbaik berdasarkan akurasi: Decision Tree
- Rangkuman hasil:
- - Logistic Regression: 0.8636
- - KNN: 0.7955
- - Naive Bayes: 0.6970
- - Decision Tree: 0.9167
- Model yang dipilih memiliki akurasi tertinggi di antara keempat model yang diuji.
- Namun, akurasi bukan satu-satunya indikator — Precision, Recall, dan F1 Score juga perlu dipertimbangkan terutama jika data tidak seimbang.
- Logistic Regression cocok jika interpretabilitas penting, sedangkan Decision Tree memberikan visualisasi pohon keputusan yang intuitif.
- KNN bergantung pada pemilihan nilai K, dan performanya dapat menurun jika data sangat besar.
- Naive Bayes bekerja baik pada data yang bersifat independen antar fitur.

Pemilihan akhir model tetap harus mempertimbangkan konteks aplikasi, kebutuhan interpretasi, serta performa menyeluruh.

THANK YOU

For your attention