



LOGISTIC REGRESSION, K- NEAREST NEIGHBOR, NÄÏVE BAYES, DECISION TREE



OLEH KELOMPOK 3

NAMA ANGGOTA



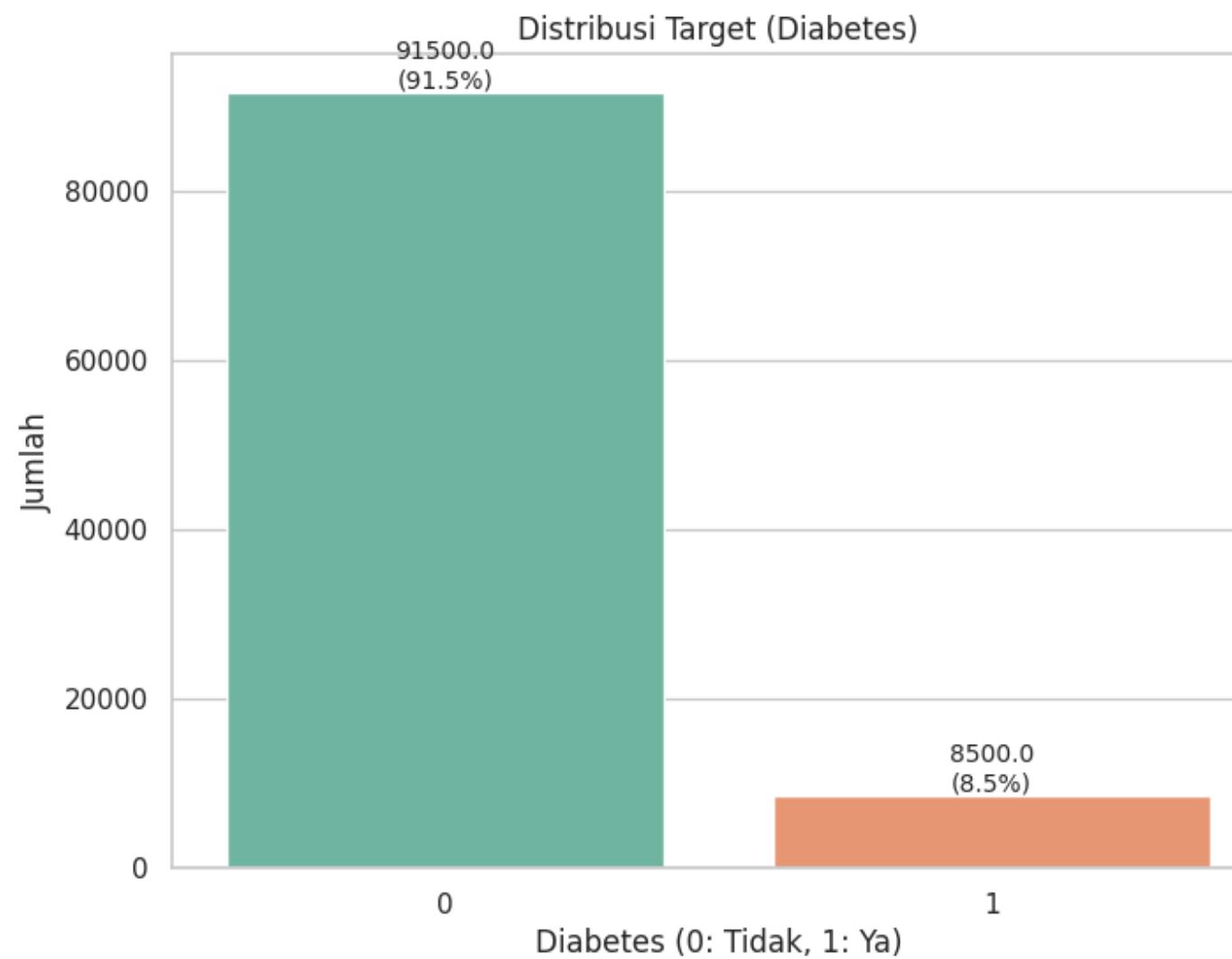
- Meutia Aini
(2208107010005)
- Akhsania Maisa Rahmah
(2208107010017)
- Fadli Ahmad Yazid
(2208107010032)
- Muhammad Mahathir
(2208107010056)
- Muhammad Aufa Zaikra
(2208107010070)

DATASET

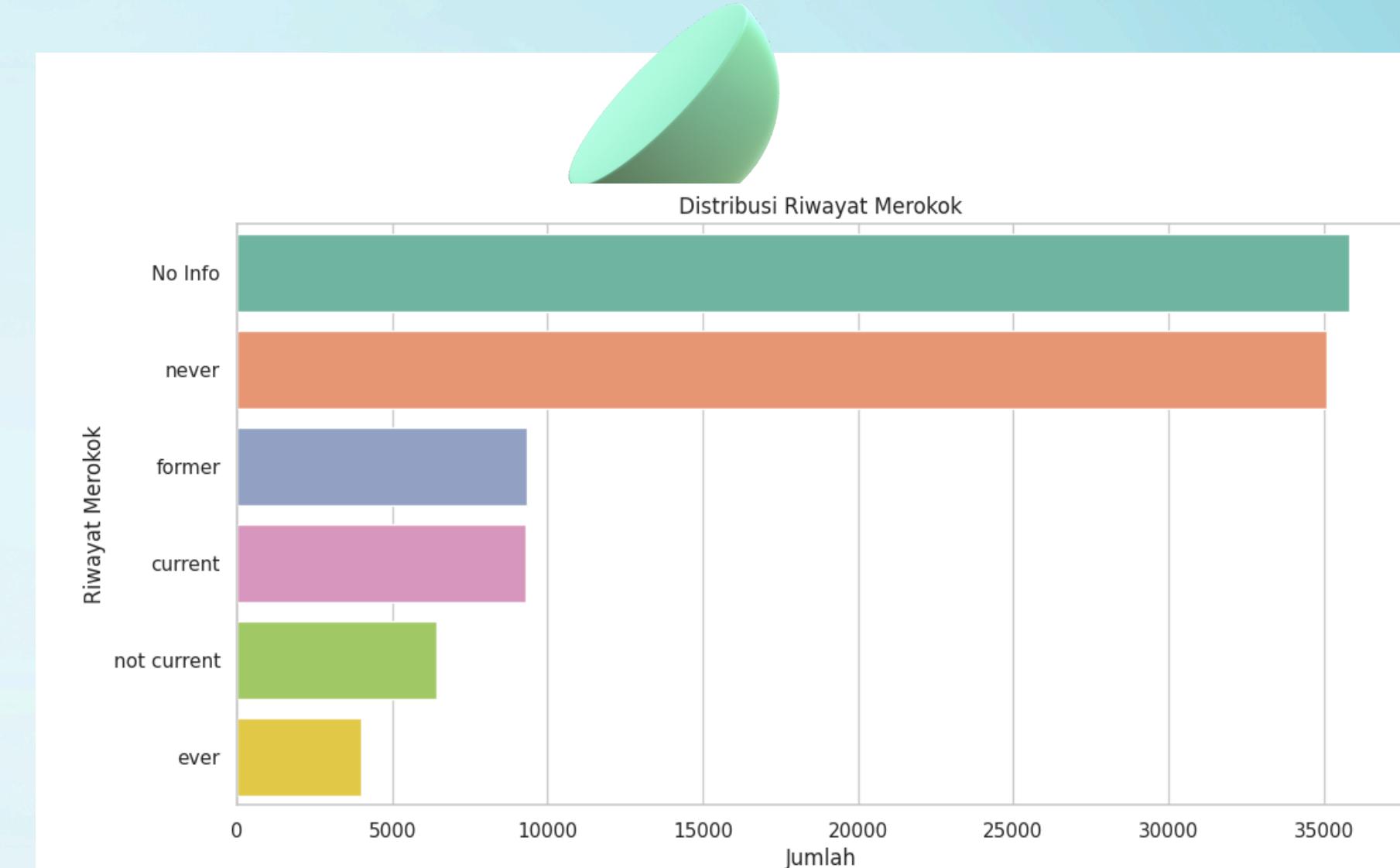
Dataset “Diabetes Prediction” berisi 100.000 data pasien dengan fitur kesehatan seperti usia, jenis kelamin, hipertensi, penyakit jantung, status merokok, BMI, hemoglobin A1c, dan glukosa darah. Target klasifikasinya adalah kolom diabetes (1 = diabetes, 0 = tidak). Dataset ini digunakan untuk membangun model klasifikasi biner menggunakan algoritma supervised learning guna memprediksi risiko diabetes.



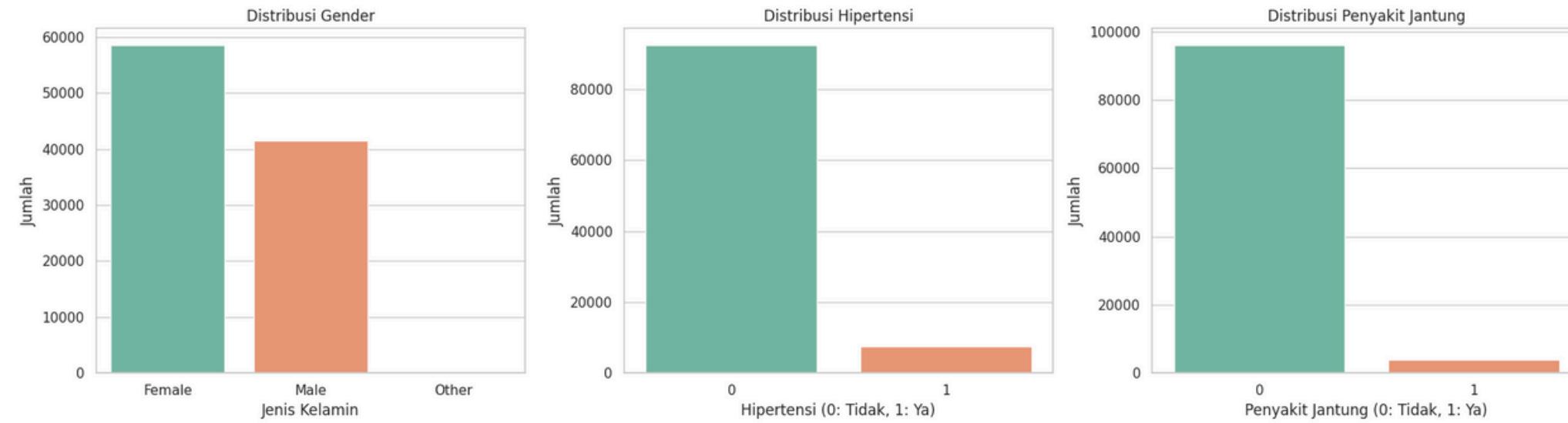
VISUALISASI AWAL DATA



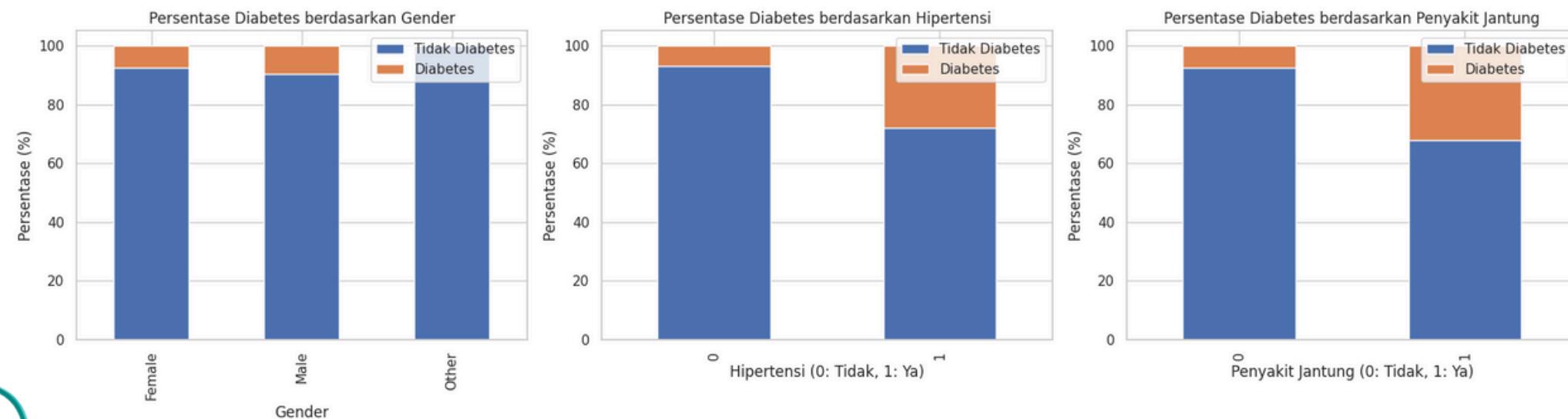
Visualisasi distribusi target menunjukkan bahwa dataset tidak seimbang, dengan jumlah data untuk kelas tidak diabetes jauh lebih banyak dibandingkan kelas diabetes.



Visualisasi riwayat merokok menunjukkan distribusi data berdasarkan kategori riwayat merokok.

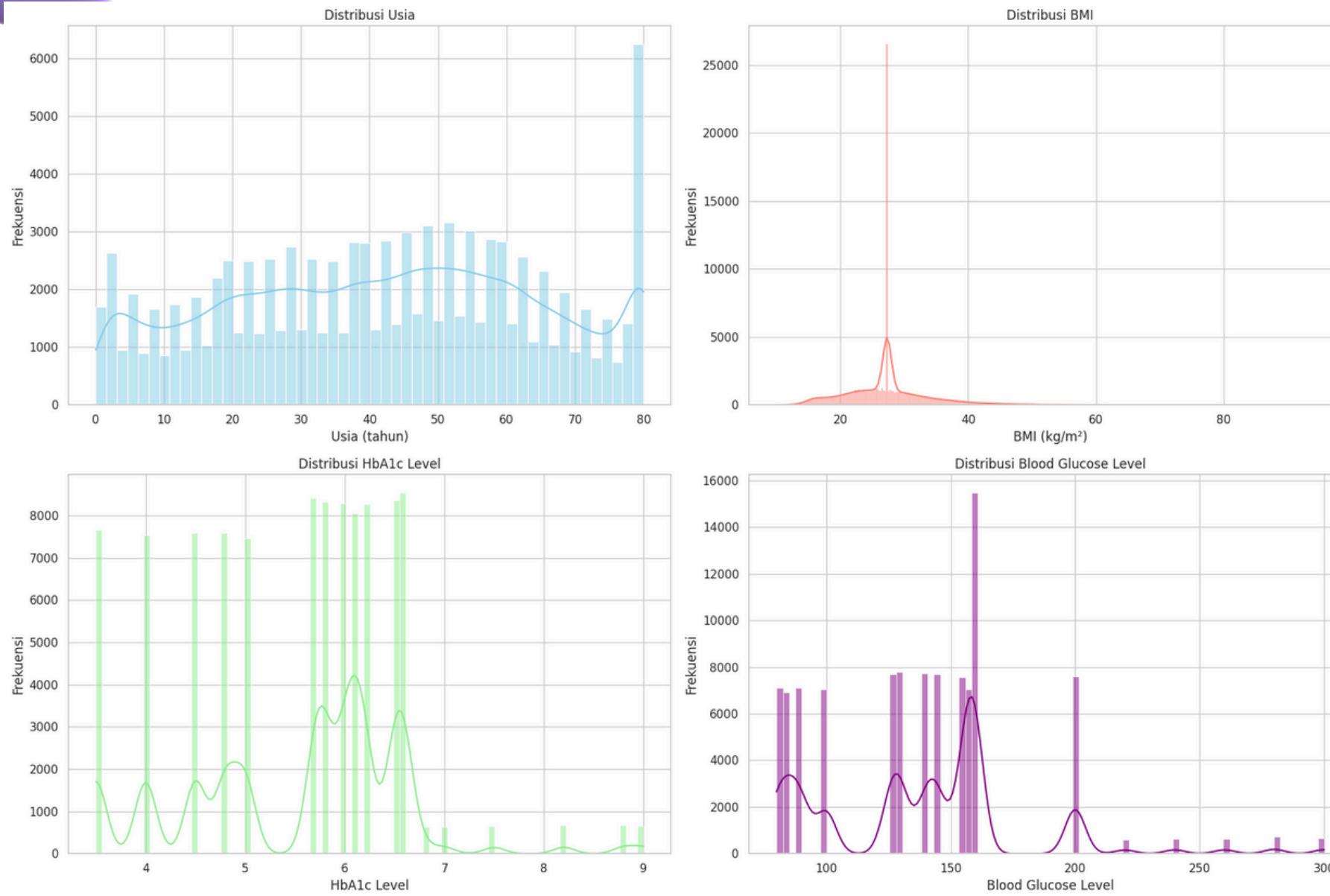


Distribusi gender menunjukkan mayoritas data adalah Female, sementara mayoritas data tidak memiliki hipertensi atau penyakit jantung. persingkat sedikit

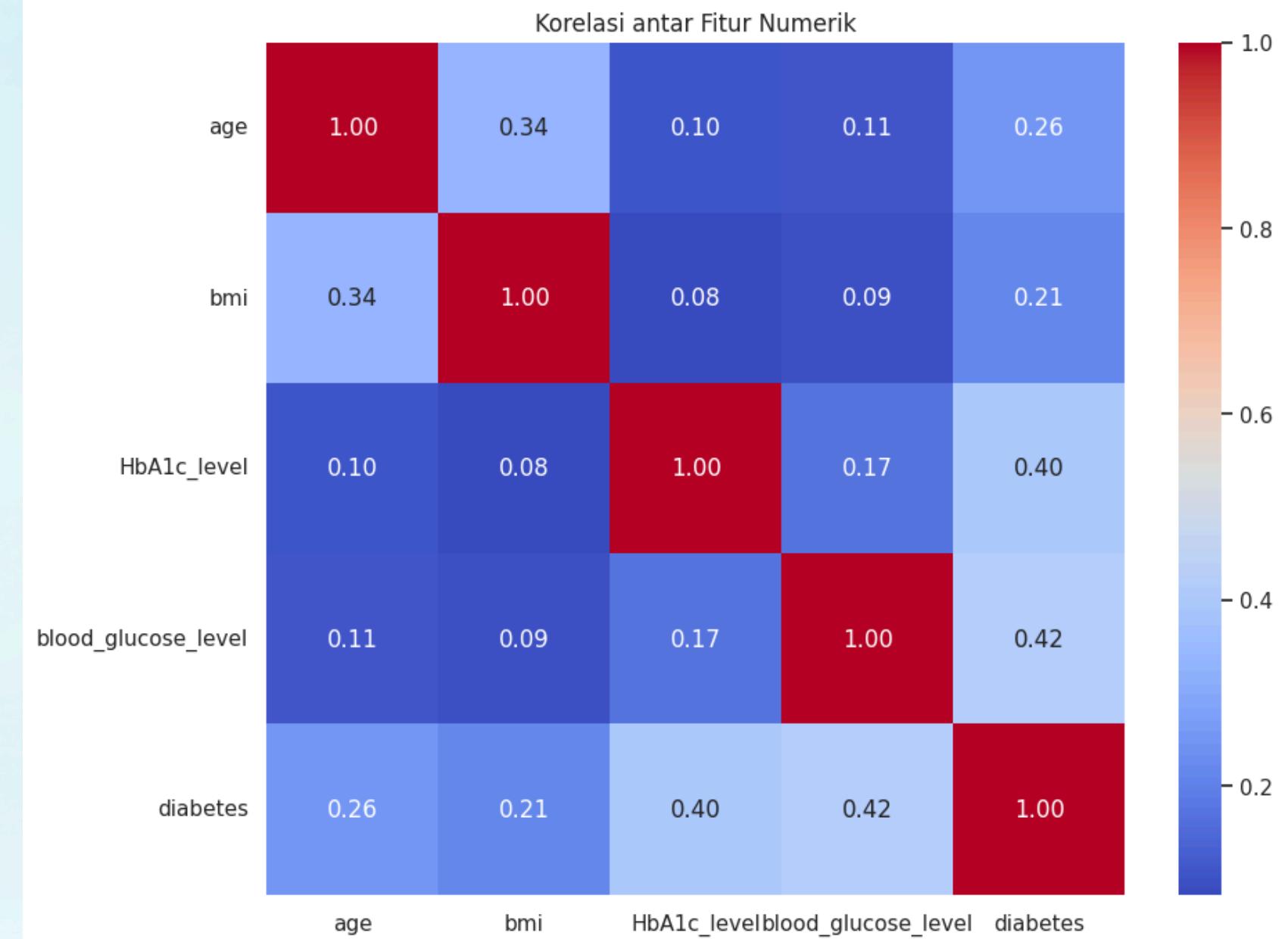


Visualisasi ini membantu kita memahami bagaimana fitur kategorikal tersebut berhubungan dengan kemungkinan seseorang menderita diabetes.

VISUALISASI AWAL DATA

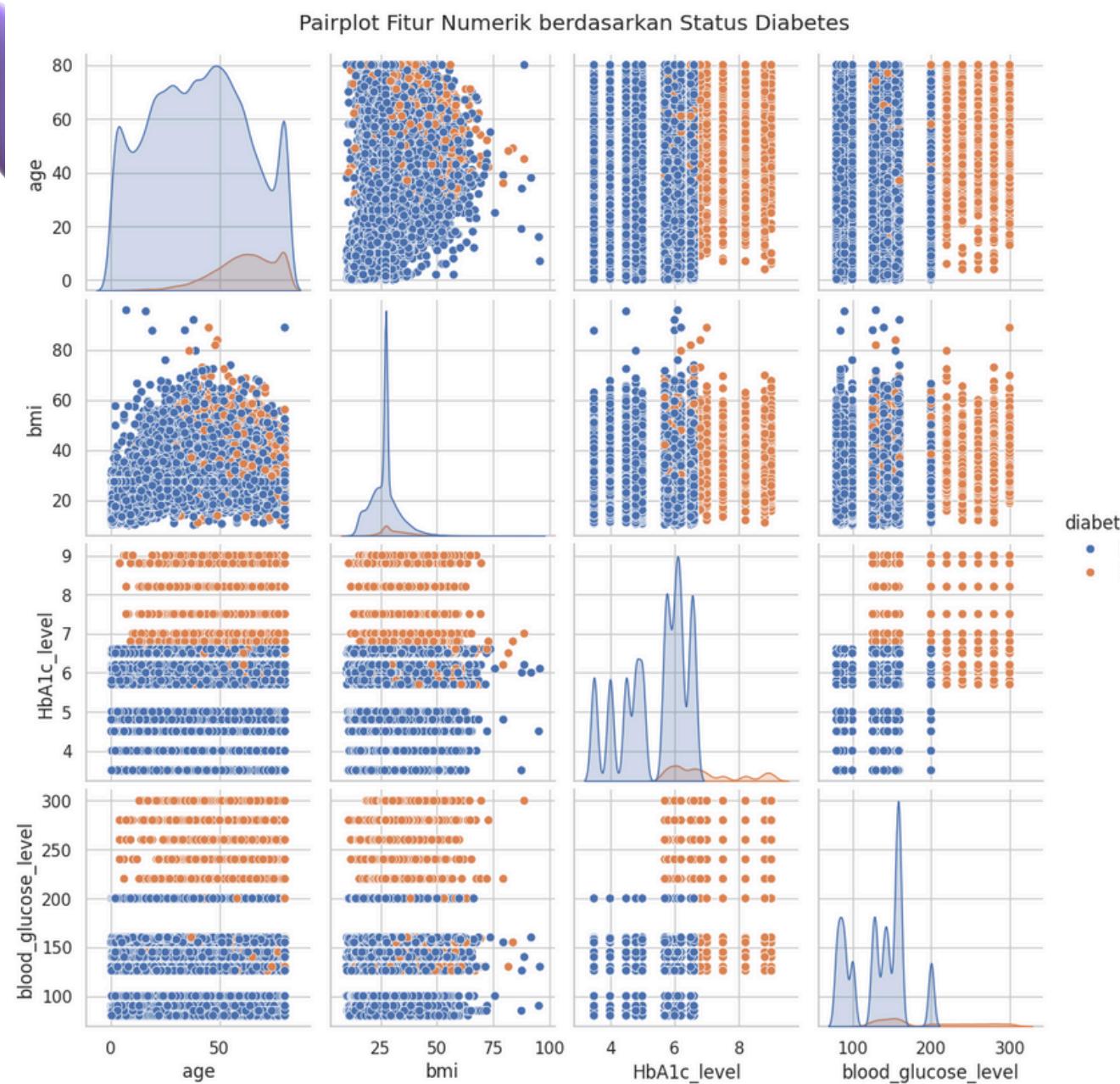


Visualisasi fitur numerik menunjukkan distribusi data untuk usia, BMI, level HbA1c, dan kadar glukosa darah. Hal ini membantu kita memahami pola distribusi dan potensi outlier dalam data.



Heatmap korelasi menunjukkan tingkat korelasi antar fitur numerik dalam dataset. Warna merah menunjukkan korelasi yang lebih kuat, sedangkan warna biru menunjukkan korelasi yang lebih lemah.

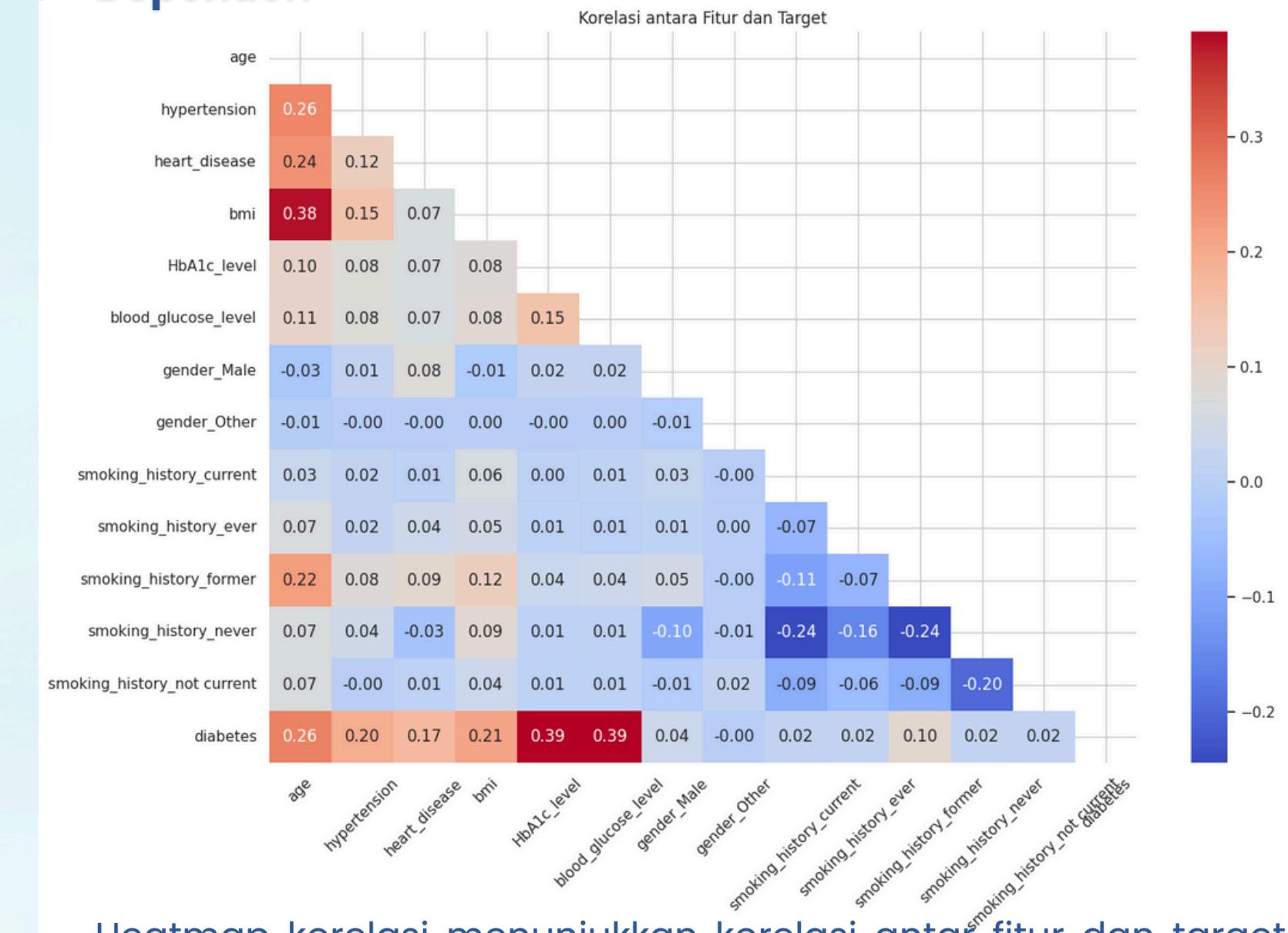
VISUALISASI AWAL



Pairplot menunjukkan hubungan antar fitur numerik berdasarkan status diabetes. Hal ini membantu kita melihat bagaimana fitur-fitur tersebut saling berkaitan dan bagaimana pola distribusi data berdasarkan status diabetes.

EKSPLORASI DATA DAN PRA-PEMROSESAN

Analisis Korelasi antara Variabel Independen dan Dependen

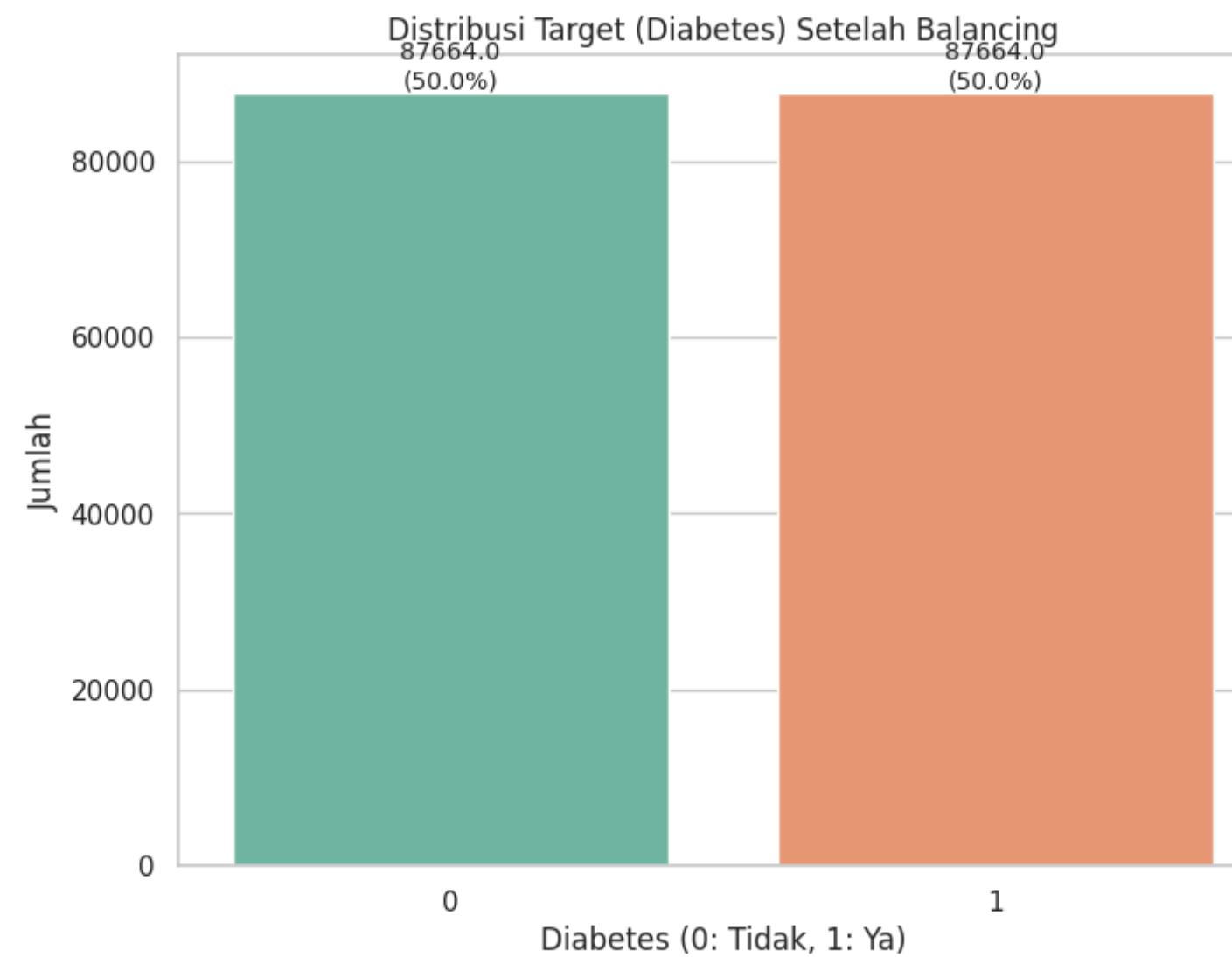


Heatmap korelasi menunjukkan korelasi antar fitur dan target. Nilai korelasi berkisar antara -1 hingga 1, di mana nilai 1 menunjukkan korelasi positif sempurna, nilai -1 menunjukkan korelasi negatif sempurna, dan nilai 0 menunjukkan tidak ada korelasi.

MENANGANI KETIDAKSEIMBANGAN KELAS

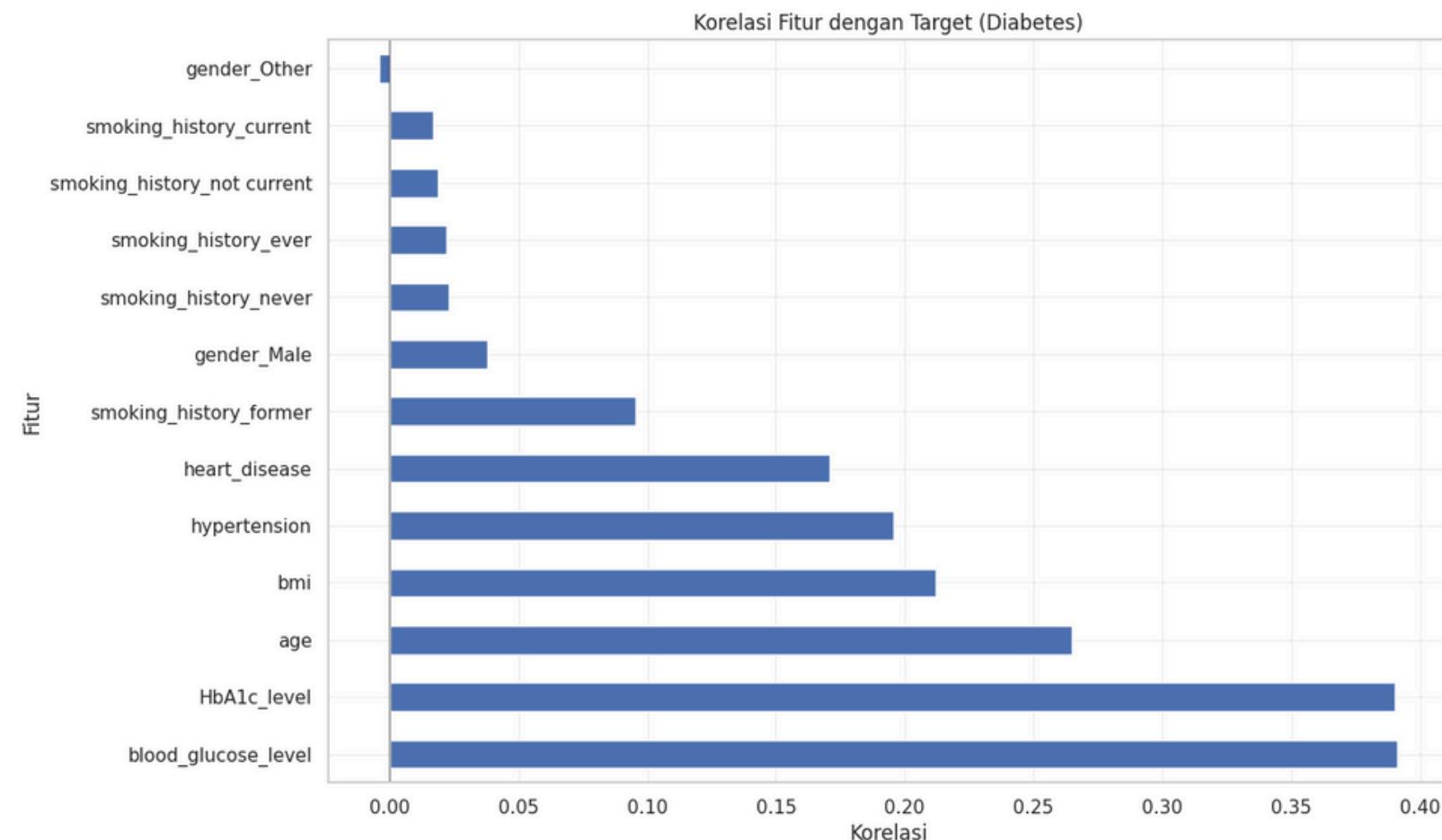
SMOTE

Sebelum balancing, kelas 0 (tidak diabetes) memiliki 87.664 data, sedangkan kelas 1 (diabetes) hanya memiliki 8.482 data. Setelah balancing dengan SMOTE, kedua kelas memiliki jumlah data yang sama, yaitu 87.664 data.



Korelasi Variabel dengan Target

Visualisasi korelasi variabel dengan target menunjukkan fitur-fitur yang memiliki korelasi tertinggi dengan diabetes. Fitur dengan korelasi positif tinggi menunjukkan bahwa peningkatan nilai fitur tersebut cenderung meningkatkan kemungkinan diabetes, sedangkan fitur dengan korelasi negatif tinggi menunjukkan sebaliknya.



IMPLEMENTASI MODEL

Train-Test Split

Pada tahap ini, kita membagi dataset menjadi data latih dan data uji dengan perbandingan 80:20.

```
Jumlah data latih: 76916
```

```
Jumlah data uji: 19230
```

```
Distribusi Target pada Data Latih:
```

```
diabetes
```

```
0 0.911774
```

```
1 0.088226
```

```
Name: proportion, dtype: float64
```

```
Distribusi Target pada Data Uji:
```

```
diabetes
```

```
0 0.911804
```

```
1 0.088196
```

```
Name: proportion, dtype: float64
```

Data dibagi menjadi 76.916 data latih dan 19.230 data uji. Parameter `stratify=y` memastikan bahwa distribusi kelas dalam data latih dan data uji sama dengan distribusi kelas dalam dataset asli.



IMPLEMENTASI MODEL

Membangun Model Logistic Regression

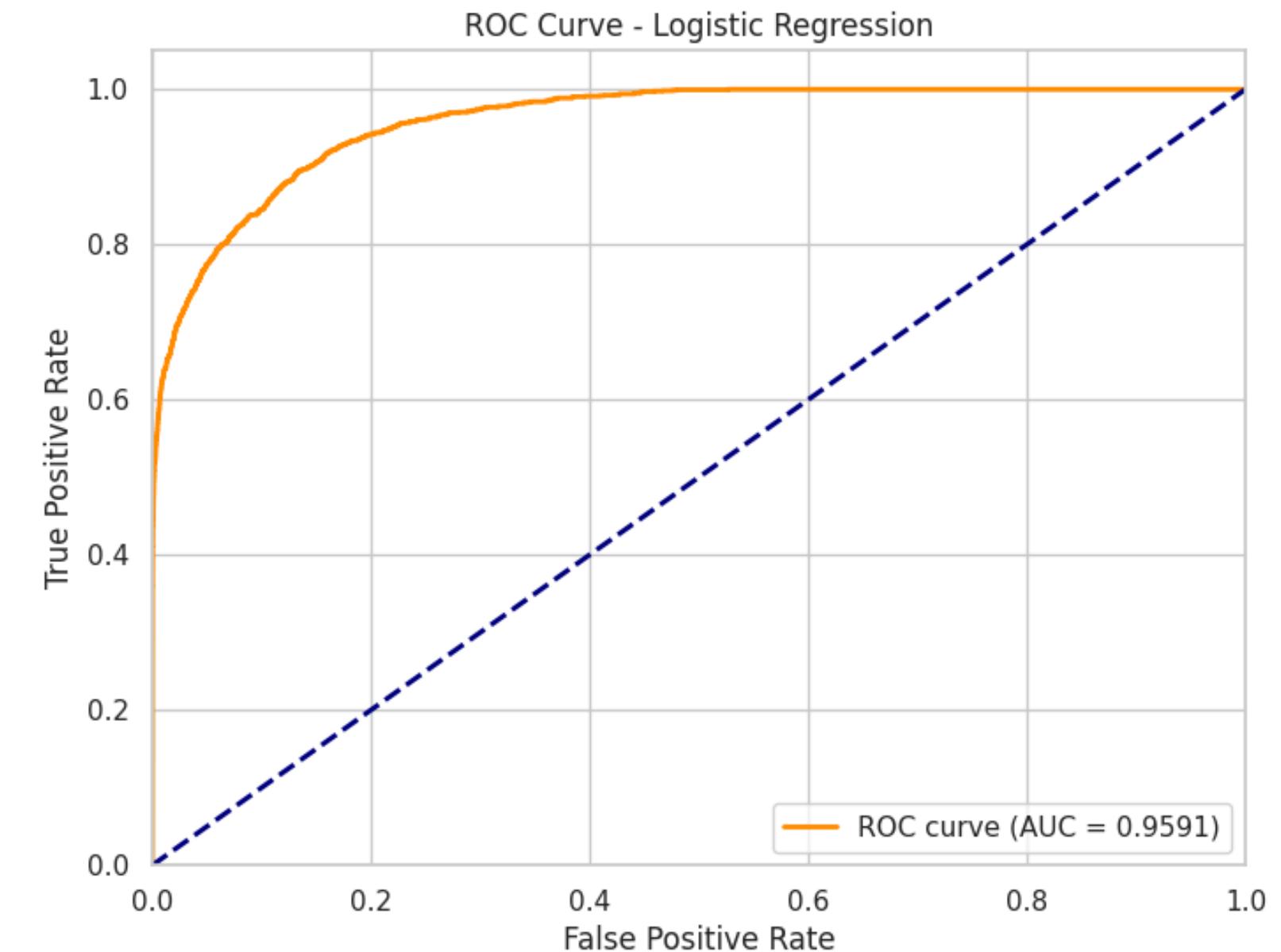
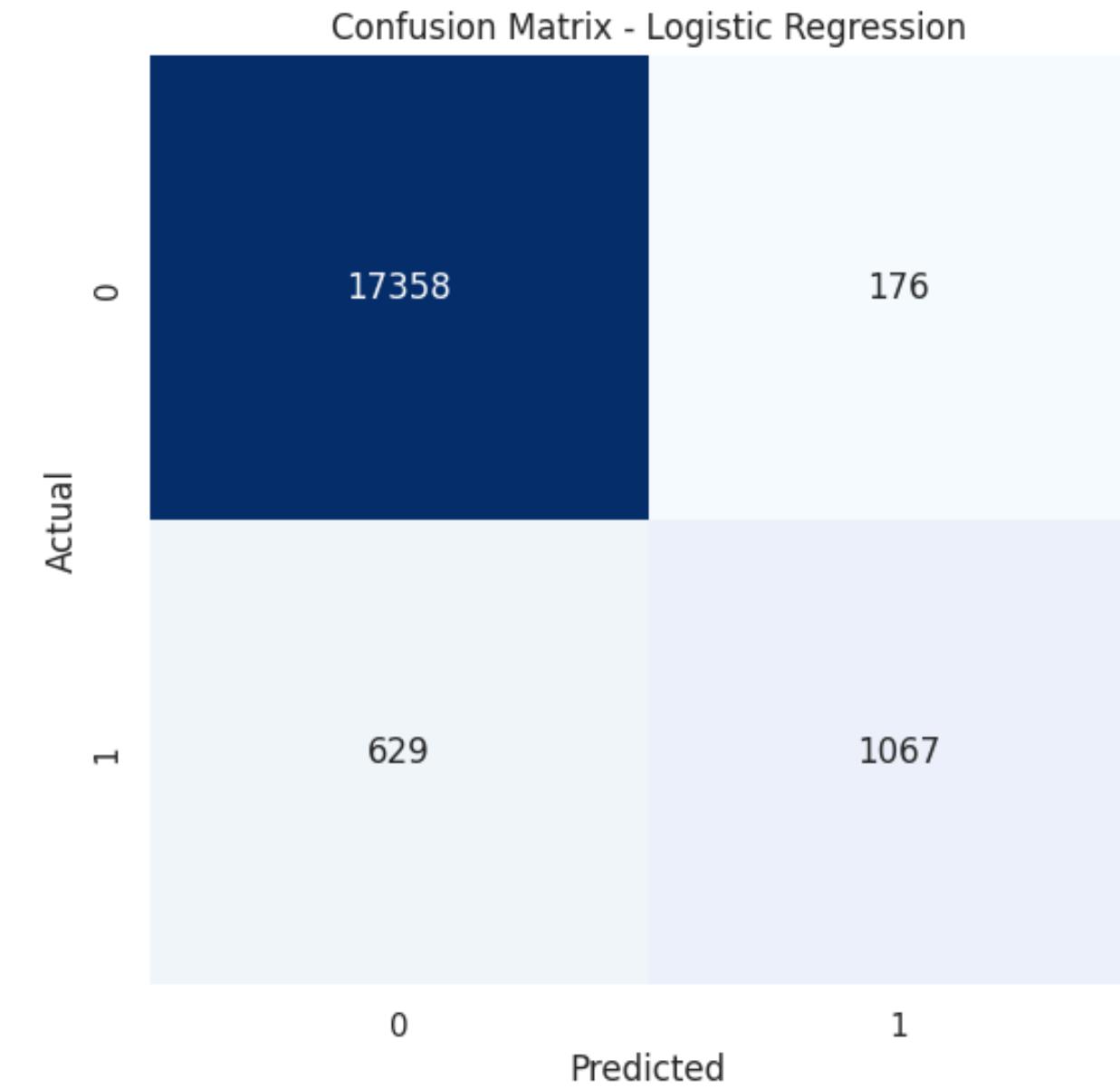
Evaluasi Model Logistic Regression:

Accuracy: 0.9581

Precision: 0.8584

Recall: 0.6291

F1 Score: 0.7261



Model Logistic Regression mencapai akurasi 95.81%, precision 85.84%, recall 62.91%, dan F1 score 72.61%. Confusion matrix dan ROC curve juga ditampilkan untuk evaluasi model yang lebih mendalam.

IMPLEMENTASI MODEL

Membangun Model K-Nearest Neighbors

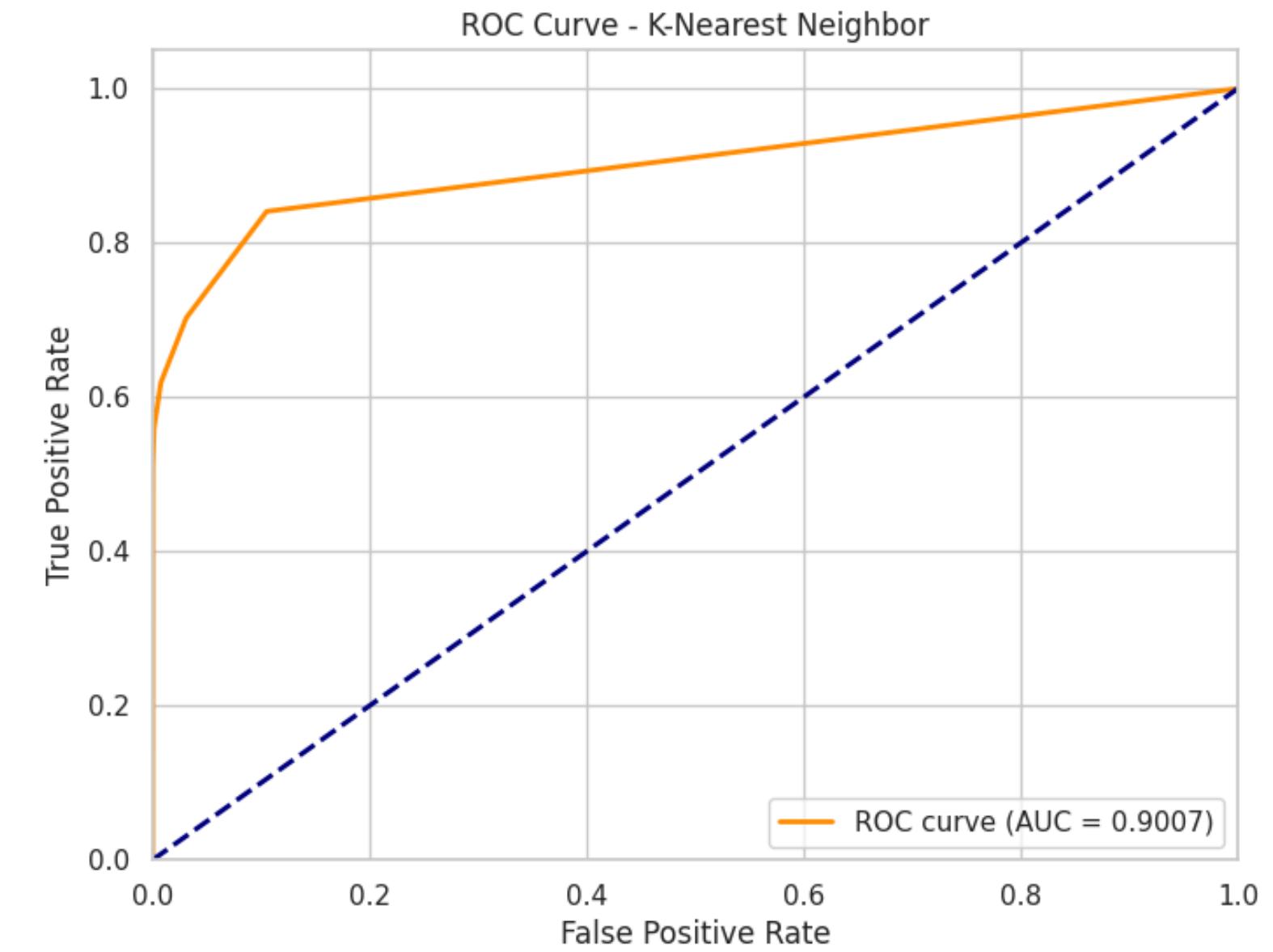
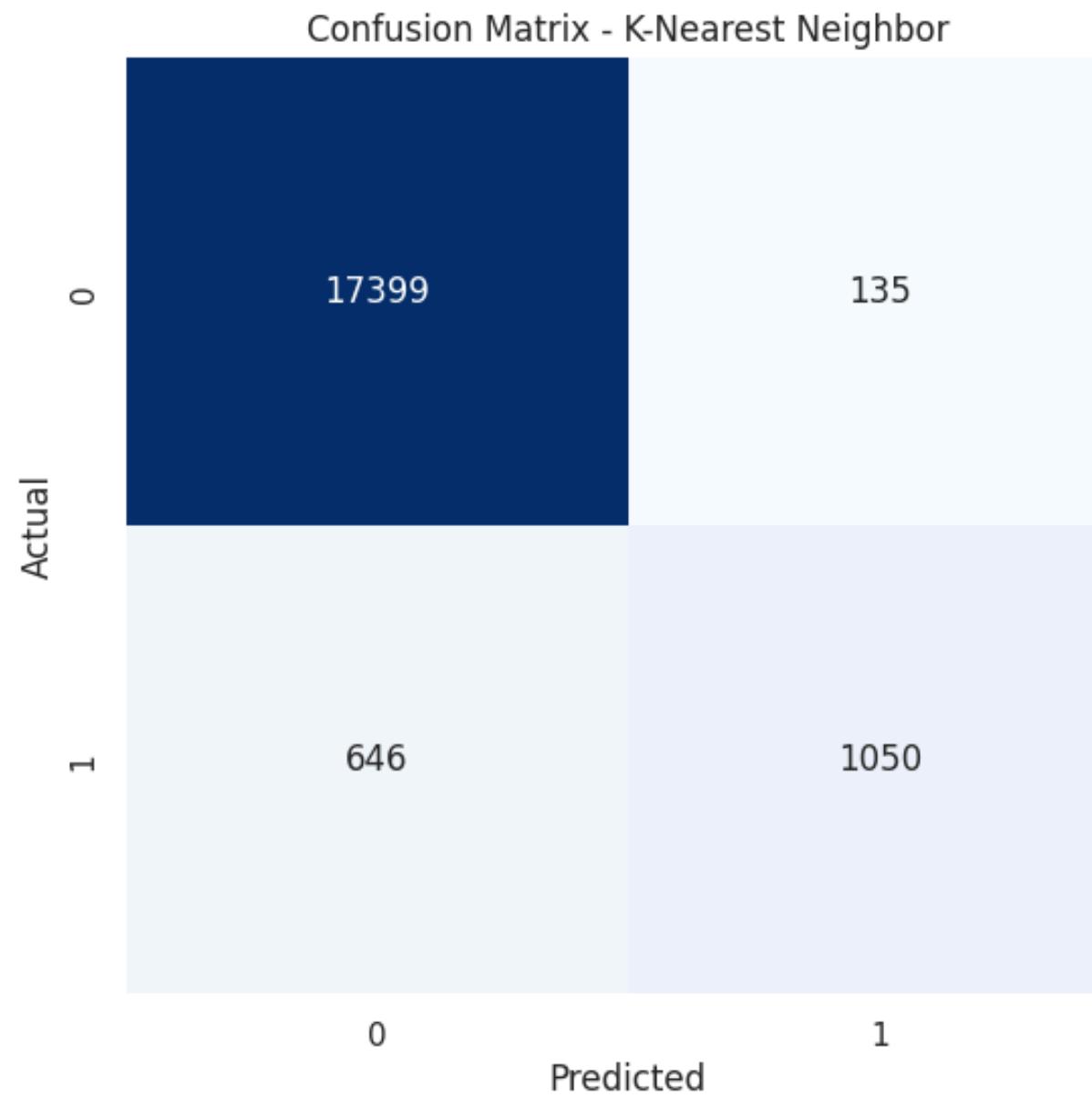
Evaluasi Model K-Nearest Neighbor:

Accuracy: 0.9594

Precision: 0.8861

Recall: 0.6191

F1 Score: 0.7289



Model K-Nearest Neighbor mencapai akurasi 95.94%, precision 88.61%, recall 61.91%, dan F1 score 72.89%. Confusion matrix dan ROC curve juga ditampilkan untuk evaluasi model yang lebih mendalam.

IMPLEMENTASI MODEL

Membangun Model Naïve Bayes

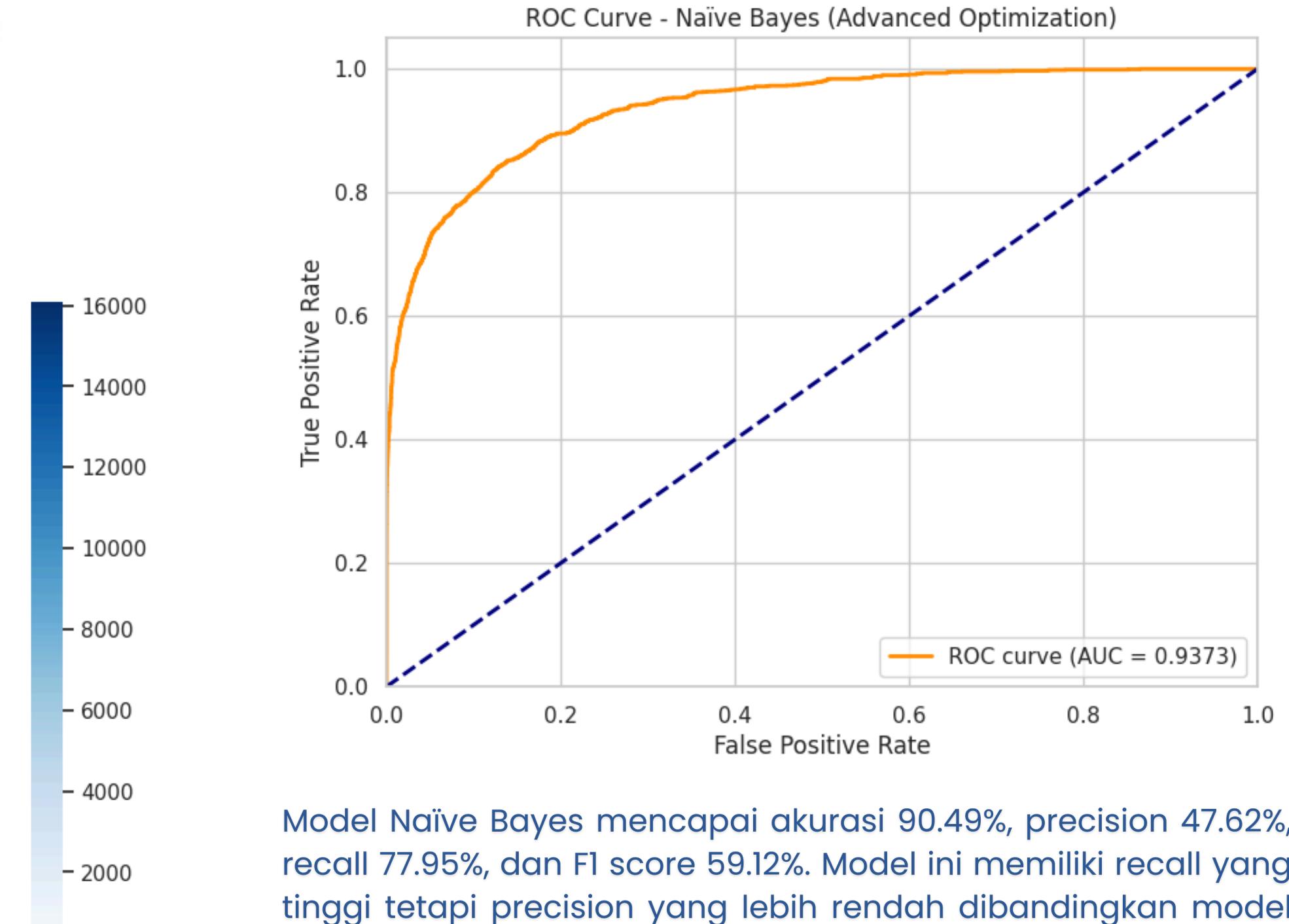
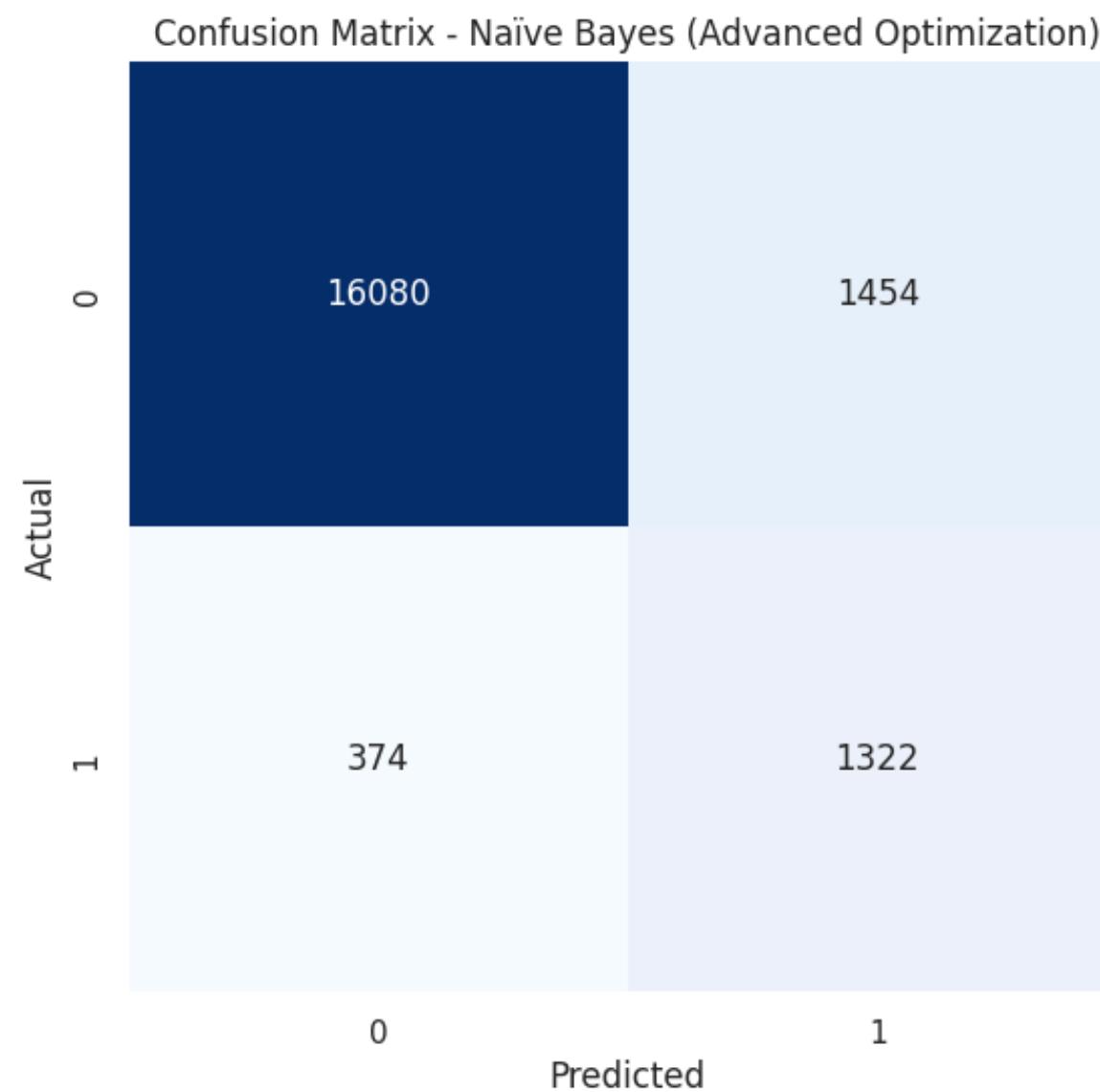
Evaluasi Model Naïve Bayes (Ensemble):

Accuracy: 0.9049

Precision: 0.4762

Recall: 0.7795

F1 Score: 0.5912



Model Naïve Bayes mencapai akurasi 90.49%, precision 47.62%, recall 77.95%, dan F1 score 59.12%. Model ini memiliki recall yang tinggi tetapi precision yang lebih rendah dibandingkan model lain.

IMPLEMENTASI MODEL

Membangun Model Decision Tree

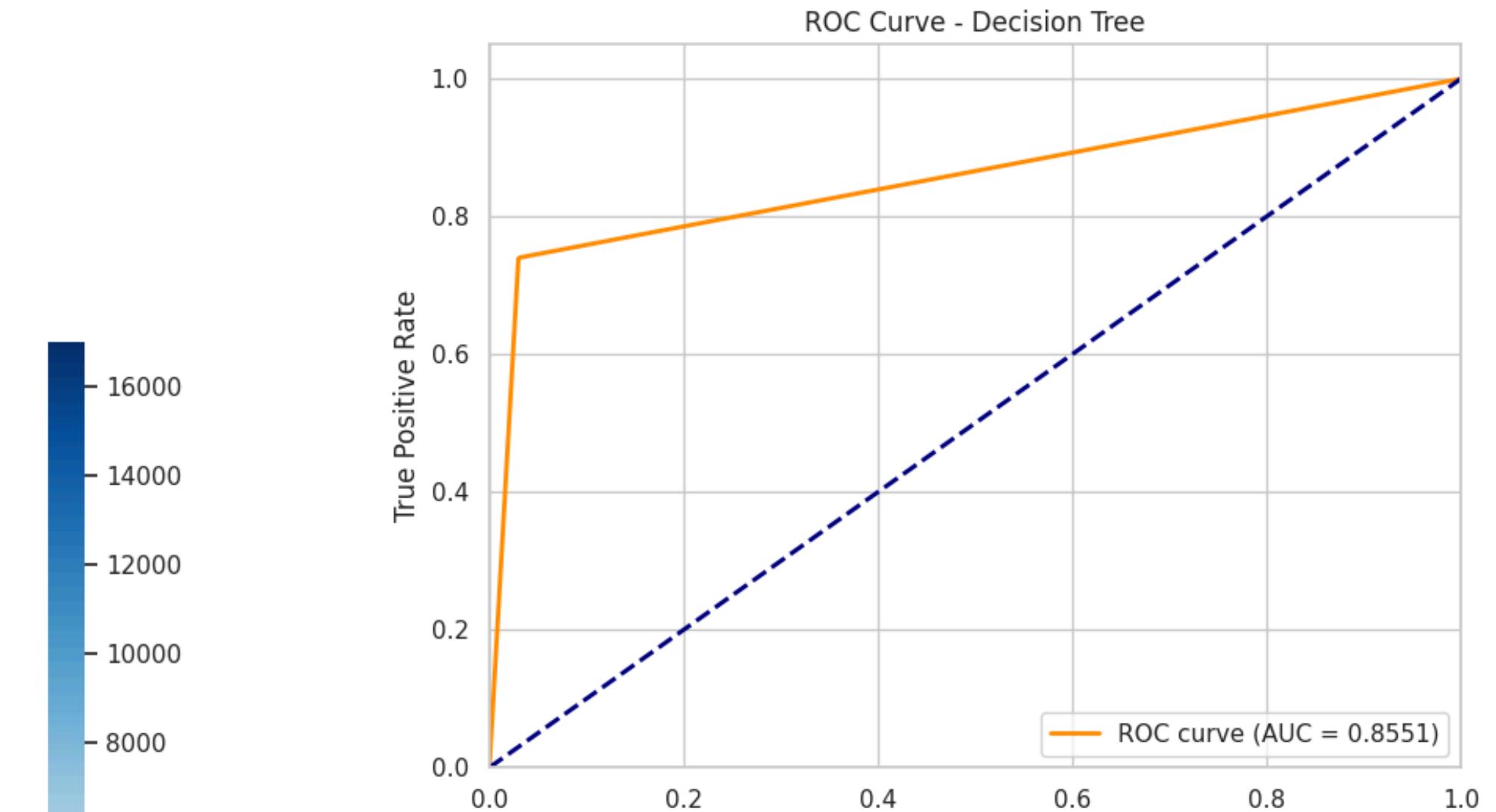
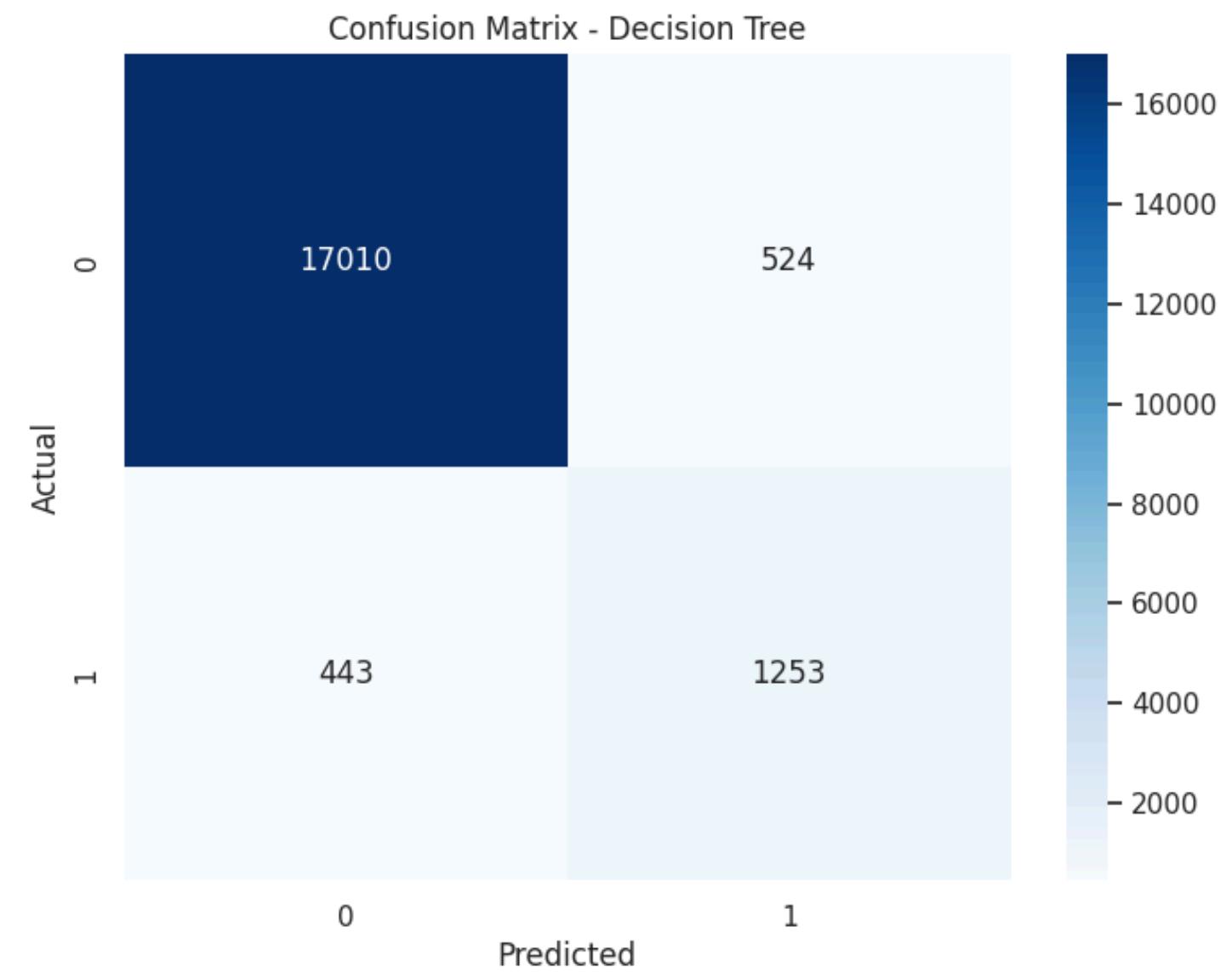
Evaluasi Model Decision Tree:

Accuracy: 0.9497

Precision: 0.7051

Recall: 0.7388

F1 Score: 0.7216



Model Decision Tree mencapai akurasi 94.97%, precision 70.51%, recall 73.88%, dan F1 score 72.16%. Model ini memiliki keseimbangan yang baik antara precision dan recall.

IMPLEMENTASI MODEL

Cross-Validation

Model: Logistic Regression

Cross-Validation Accuracy: 0.9578 ± 0.0014

All Scores: [0.95585023 0.95902023 0.956472 0.95818815 0.95922825]

Model: K-Nearest Neighbor

Cross-Validation Accuracy: 0.9591 ± 0.0013

All Scores: [0.95855434 0.95886422 0.95720006 0.95969629 0.96110042]

Model: Naïve Bayes

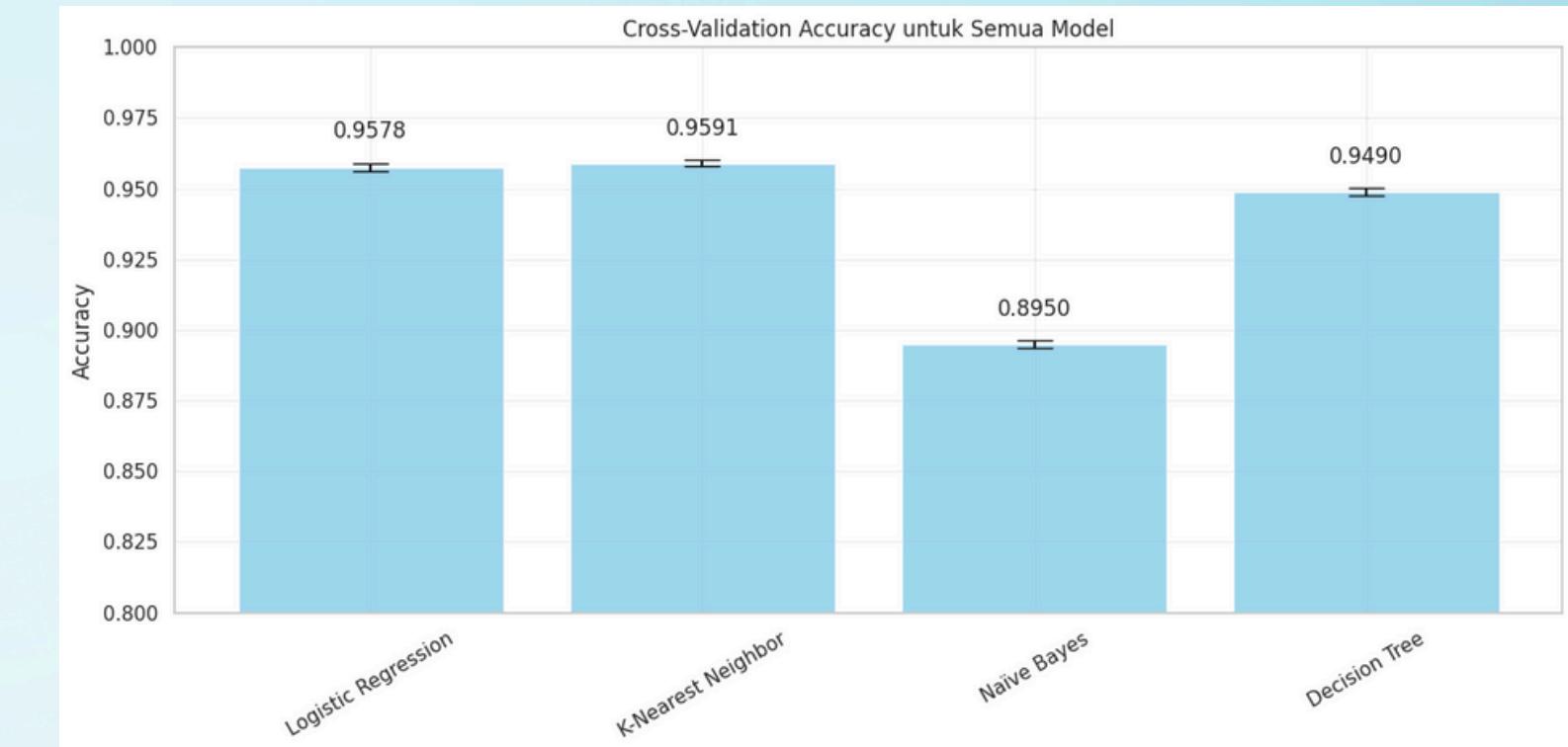
Cross-Validation Accuracy: 0.8950 ± 0.0012

All Scores: [0.89360374 0.89500234 0.89588642 0.8967705 0.89385823]

Model: Decision Tree

Cross-Validation Accuracy: 0.9490 ± 0.0013

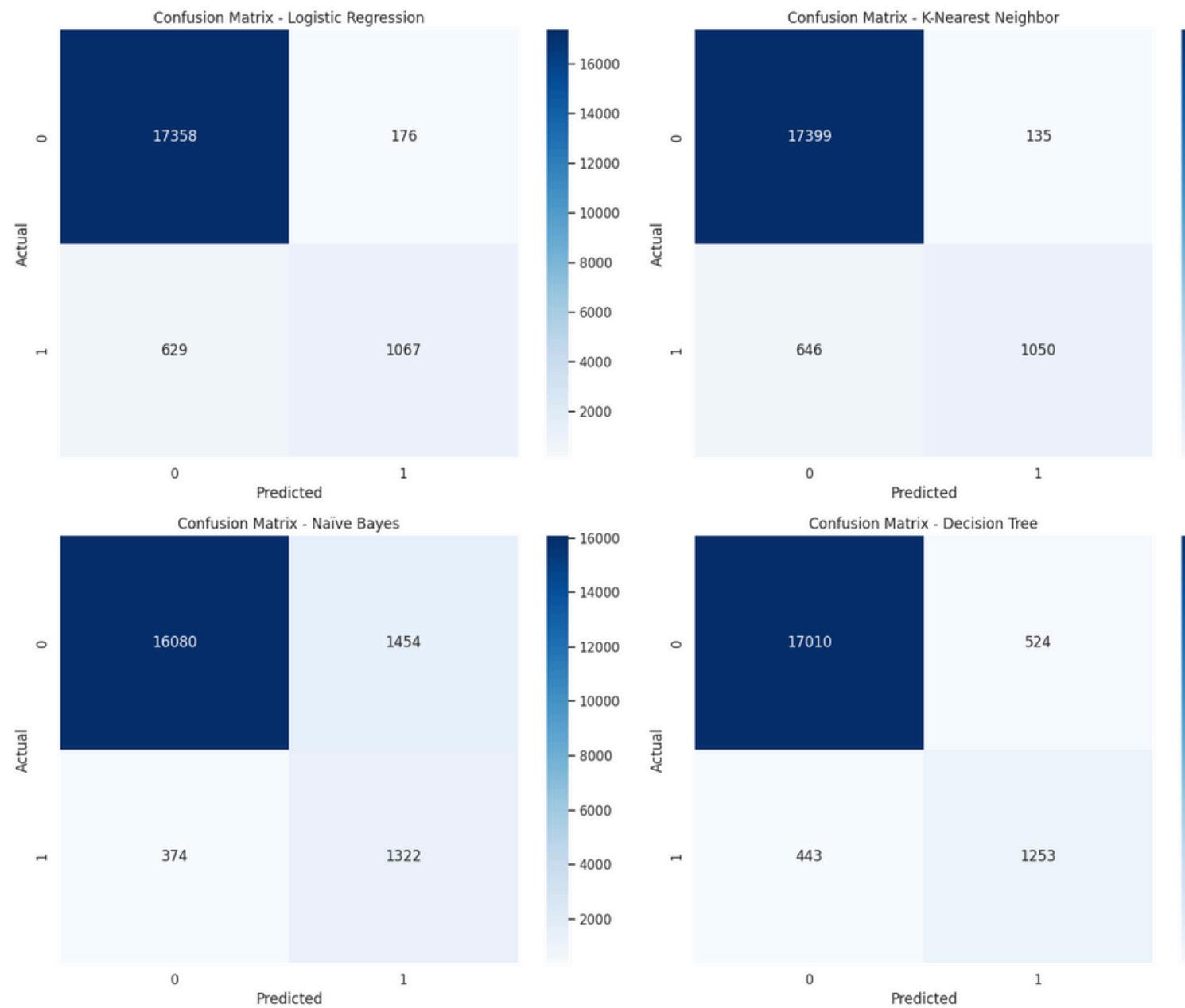
All Scores: [0.94706188 0.94804722 0.94945135 0.95054345 0.9500234]



Hasil cross-validation menunjukkan performa model pada berbagai subset data. Model K-Nearest Neighbor memiliki akurasi cross-validation tertinggi, diikuti oleh Logistic Regression, Decision Tree, dan Naïve Bayes.

EVALUASI MODEL

Confusion Matrix

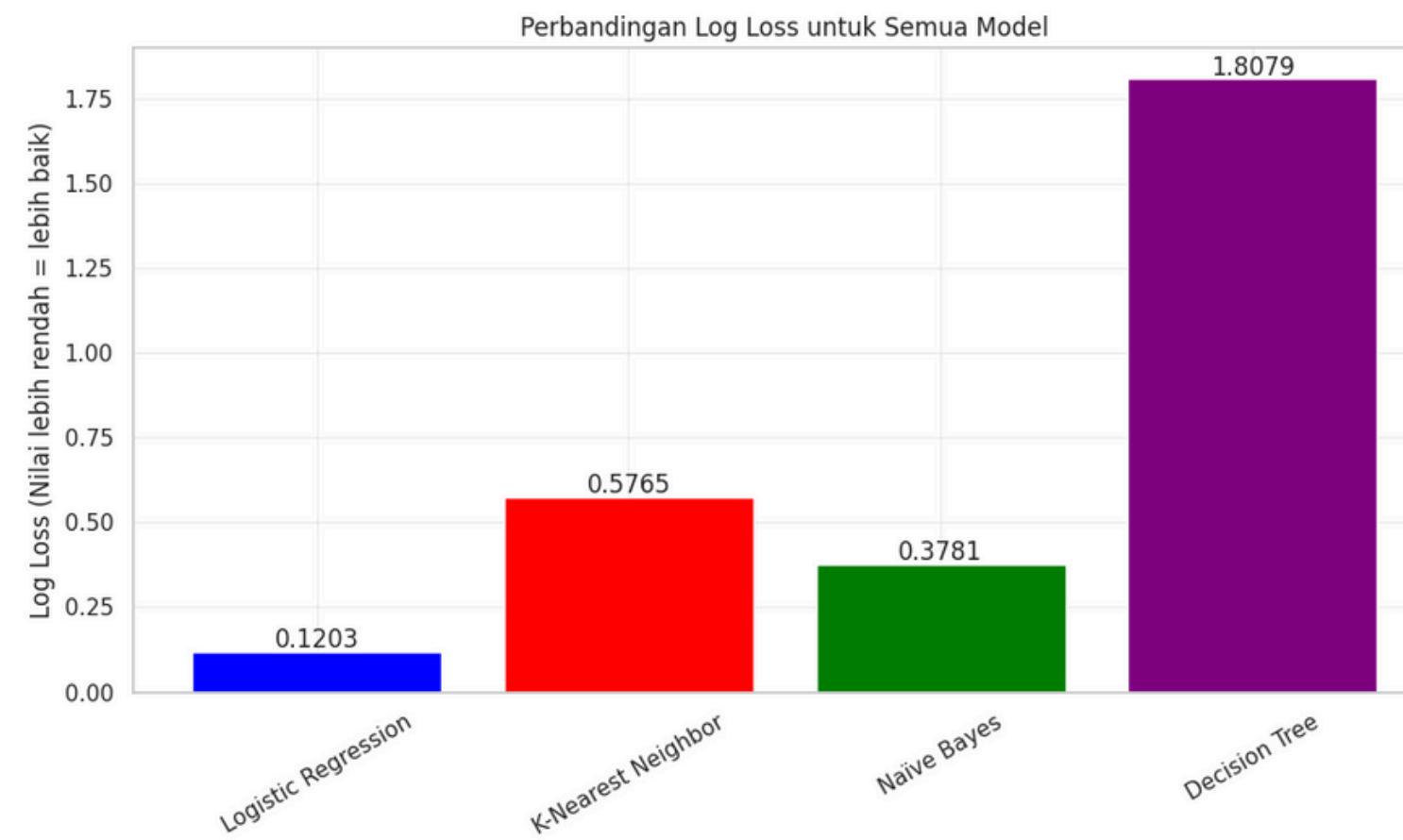


Visualisasi confusion matrix untuk semua model membantu kita membandingkan performa model dalam hal True Positive, True Negative, False Positive, dan False Negative. Dari confusion matrix, kita dapat melihat bahwa:

1. Model Logistic Regression memiliki True Negative yang tinggi tetapi False Negative yang juga relatif tinggi.
2. Model K-Nearest Neighbor memiliki pola yang mirip dengan Logistic Regression.
3. Model Naïve Bayes memiliki False Positive yang tinggi, tetapi False Negative yang rendah.
4. Model Decision Tree memiliki keseimbangan yang cukup baik antara False Positive dan False Negative.

EVALUASI MODEL

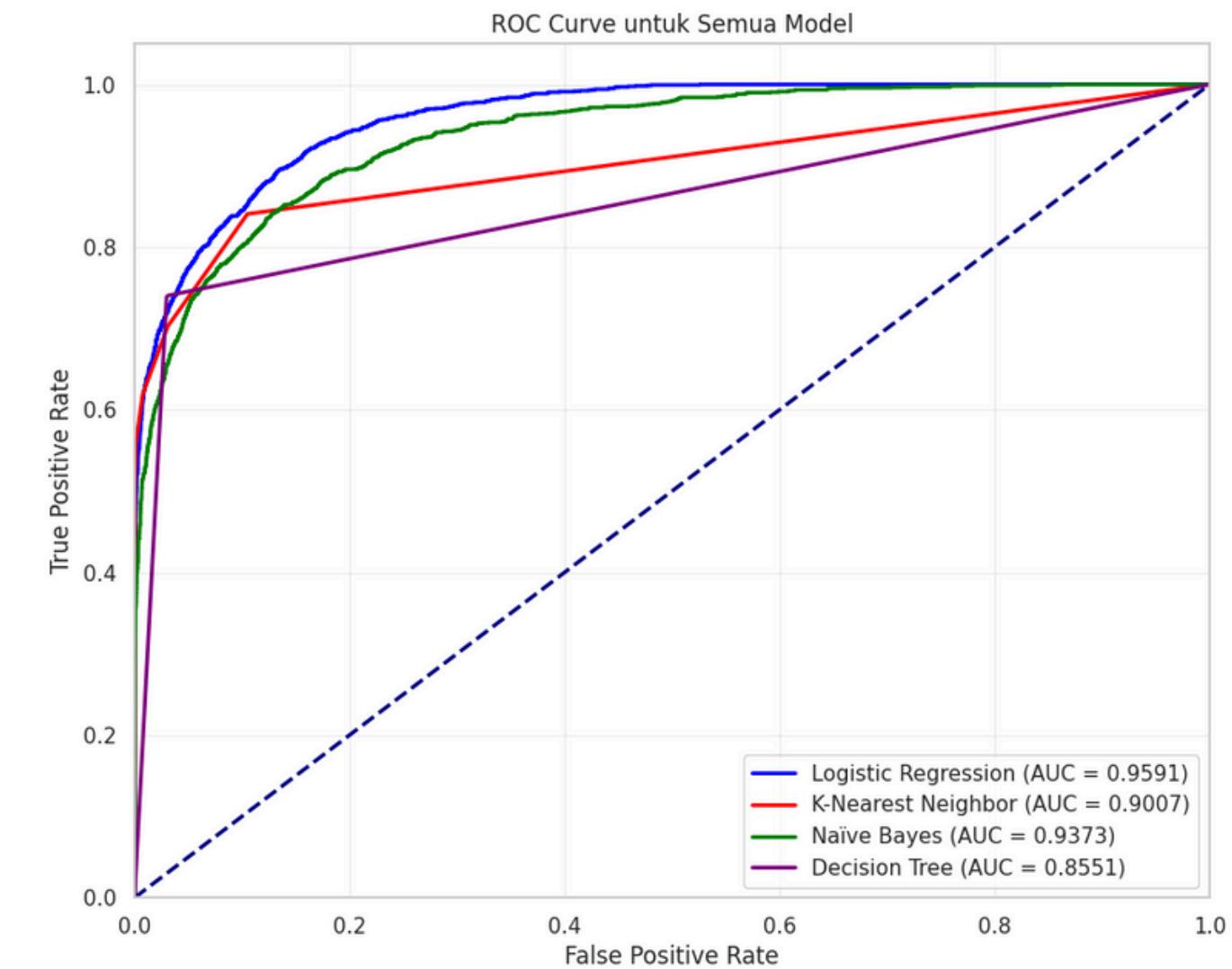
Loss Value



Dari visualisasi log loss, kita dapat melihat bahwa:

1. Model Logistic Regression memiliki log loss terendah (0.1203).
2. Model Naïve Bayes memiliki log loss kedua terendah (0.3781).
3. Model K-Nearest Neighbor memiliki log loss ketiga terendah (0.5765).
4. Model Decision Tree memiliki log loss tertinggi (1.8079).

ROC Curve



Dari ROC curve, kita dapat melihat bahwa:

1. Model Logistic Regression memiliki AUC tertinggi (0.9591).
2. Model Naïve Bayes memiliki AUC kedua tertinggi (0.9373).
3. Model K-Nearest Neighbor memiliki AUC ketiga tertinggi (0.9007).
4. Model Decision Tree memiliki AUC terendah (0.8551).

Perbandingan Metrik Evaluasi

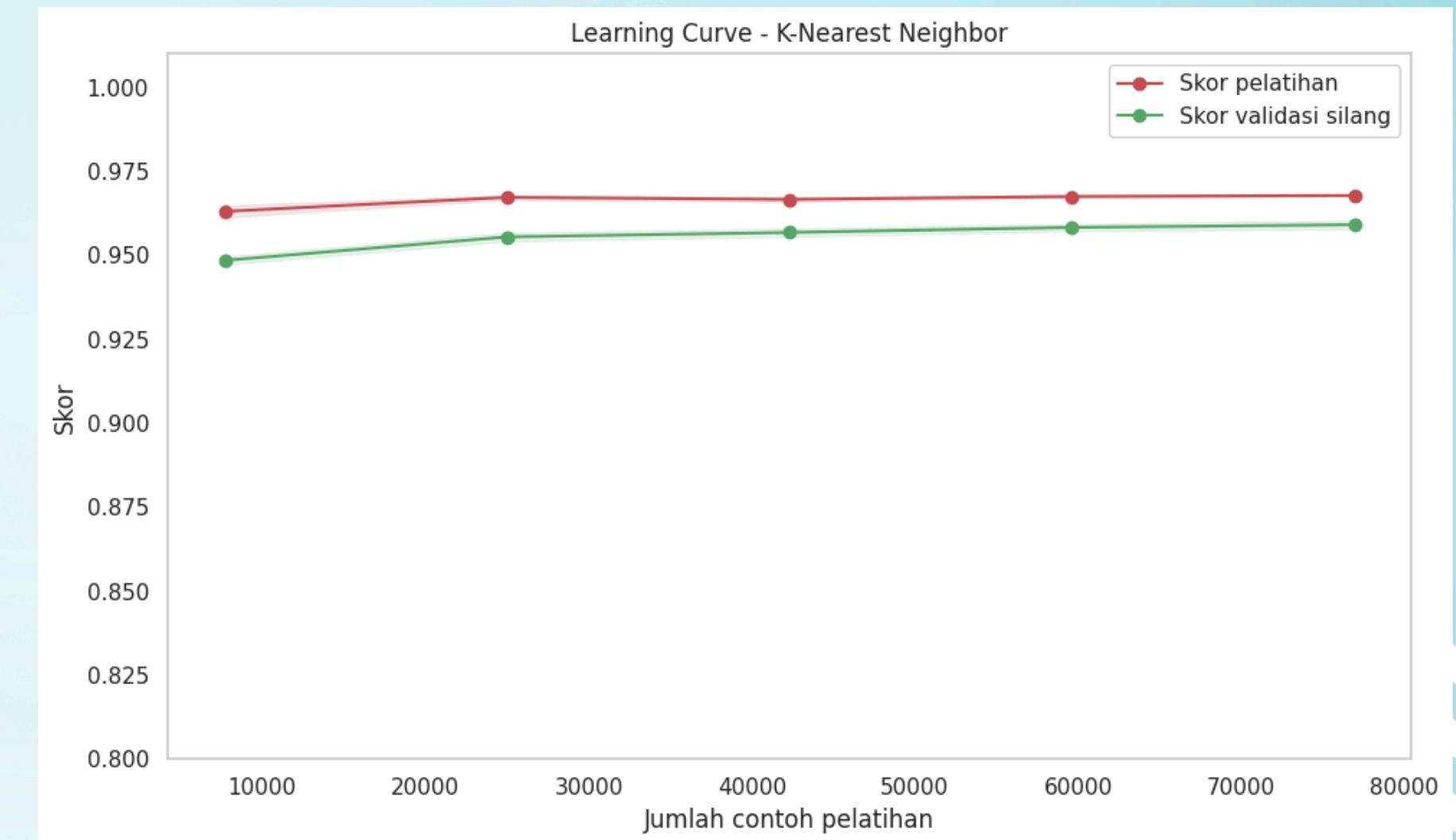
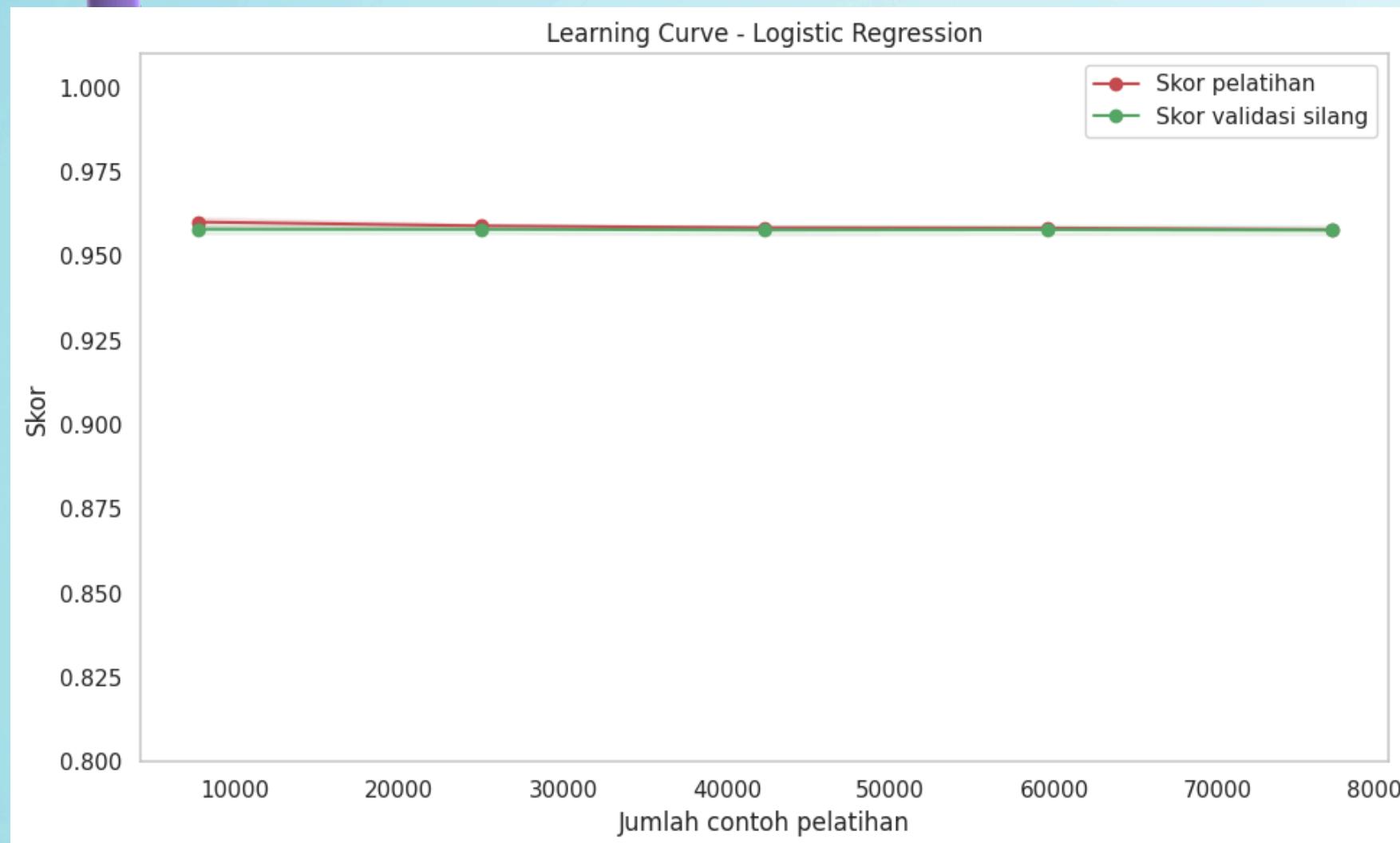
	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.958138	0.858407	0.629127	0.726097	0.959067
K-Nearest Neighbor	0.959386	0.886076	0.619104	0.728914	0.900651
Naïve Bayes	0.904940	0.476225	0.779481	0.591234	0.937264
Decision Tree	0.949714	0.705121	0.738797	0.721566	0.855113

Dari visualisasi ini, kita dapat melihat bahwa:

1. Model K-Nearest Neighbor memiliki akurasi tertinggi.
2. Model K-Nearest Neighbor juga memiliki precision tertinggi.
3. Model Naïve Bayes memiliki recall tertinggi.
4. Model K-Nearest Neighbor memiliki F1 score tertinggi.
5. Model Logistic Regression memiliki AUC tertinggi.

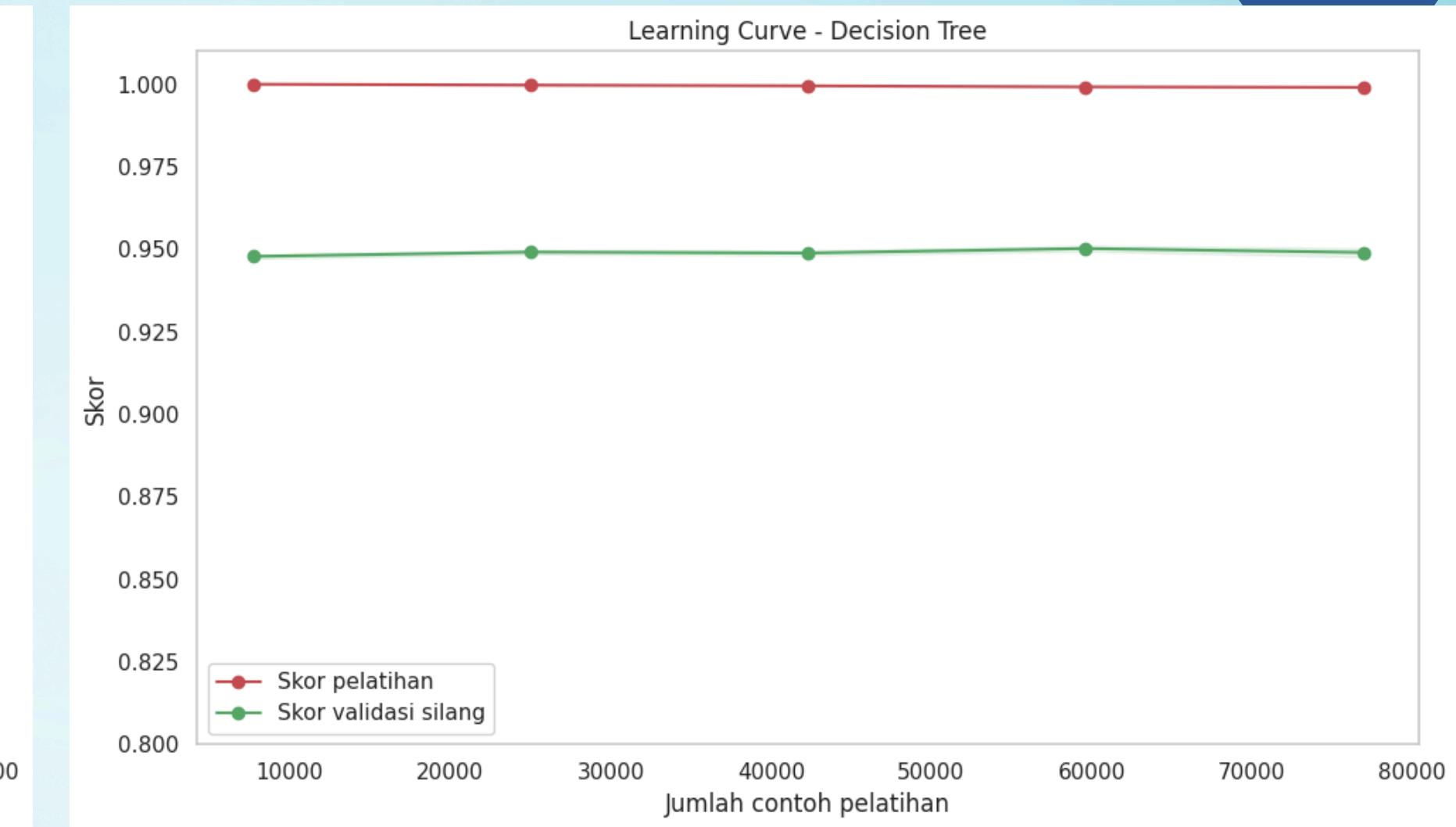
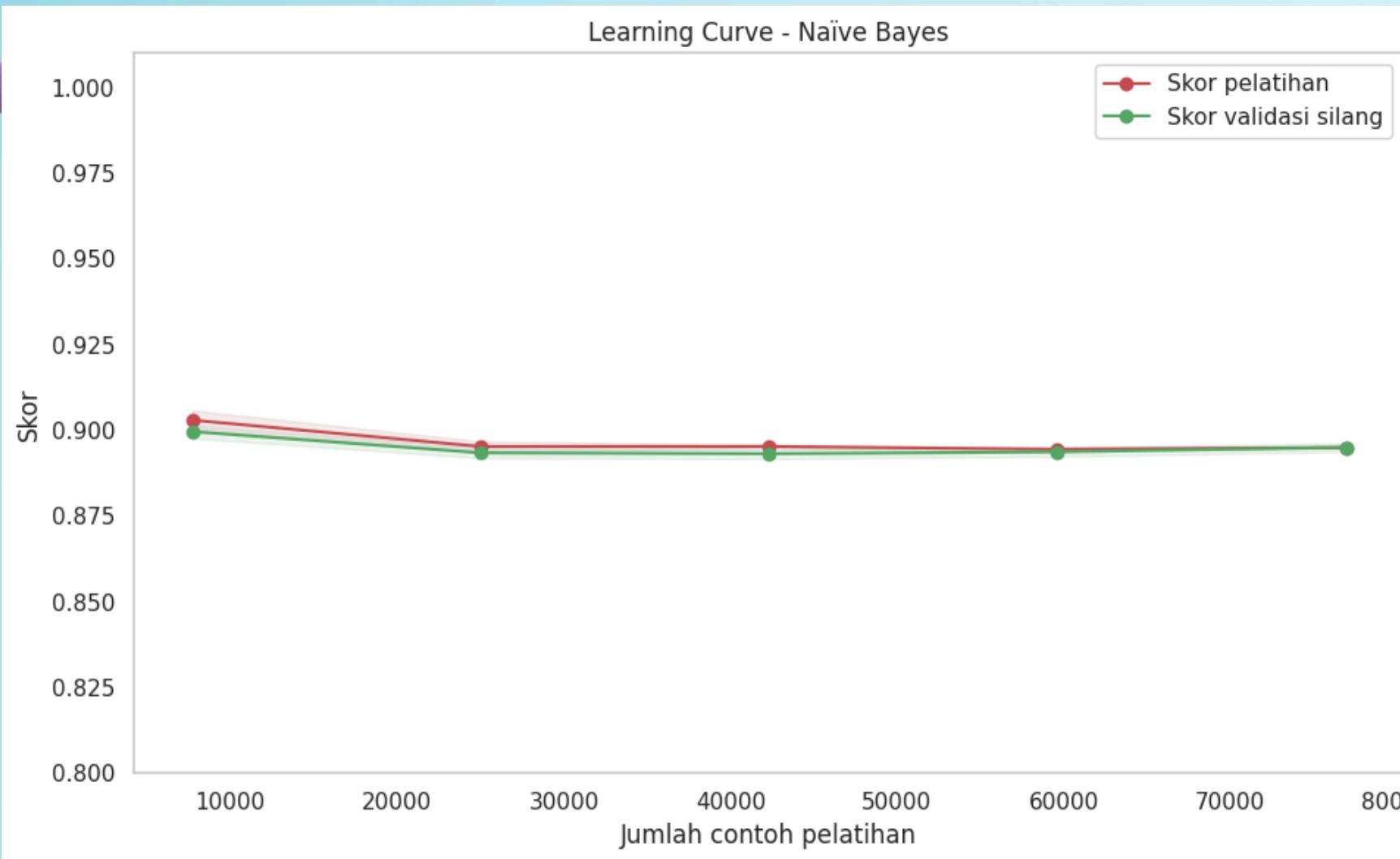
ANALISIS HASIL

Grafik Fit untuk Beberapa Algoritma



ANALISIS HASIL

Grafik Fit untuk Beberapa Algoritma



Dari learning curve, kita dapat melihat bahwa:

1. Model Logistic Regression memiliki celah kecil antara skor pelatihan dan skor validasi silang, menunjukkan bahwa model ini memiliki bias dan varians yang seimbang.
2. Model K-Nearest Neighbor memiliki celah yang sedikit lebih besar, menunjukkan bahwa model ini mungkin sedikit overfitting.
3. Model Naïve Bayes memiliki celah yang cukup besar, menunjukkan bahwa model ini mungkin underfitting.
4. Model Decision Tree memiliki celah yang paling besar, menunjukkan bahwa model ini mungkin overfitting.

TUNING HYPERPARAMETER UNTUK MODEL TERBAIK

Model terbaik berdasarkan akurasi: K-Nearest Neighbor dengan akurasi 0.9594

Hyperparameter terbaik: {'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'uniform'}

Skor terbaik: 0.9598

Evaluasi Model Terbaik Setelah Tuning:

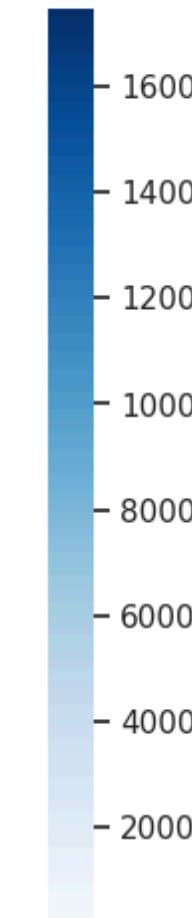
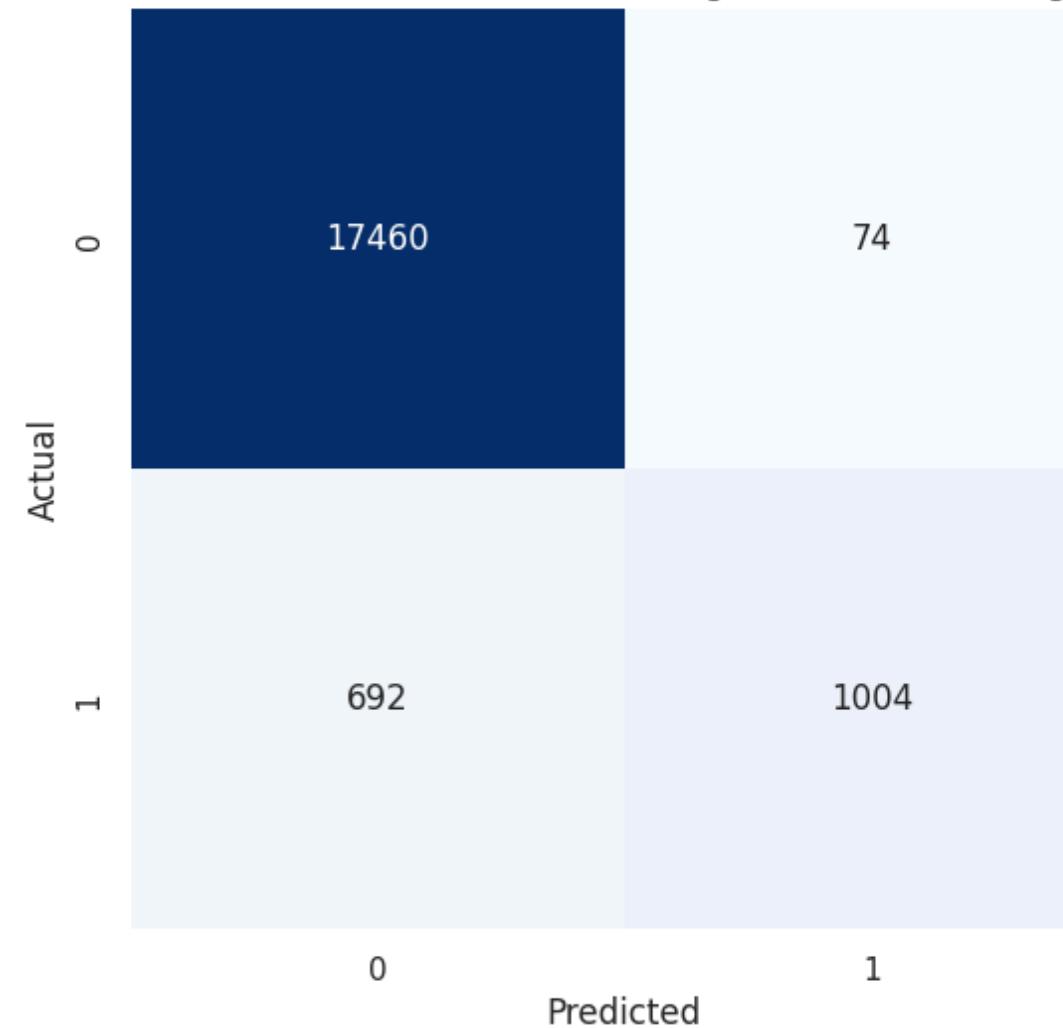
Accuracy: 0.9602

Precision: 0.9314

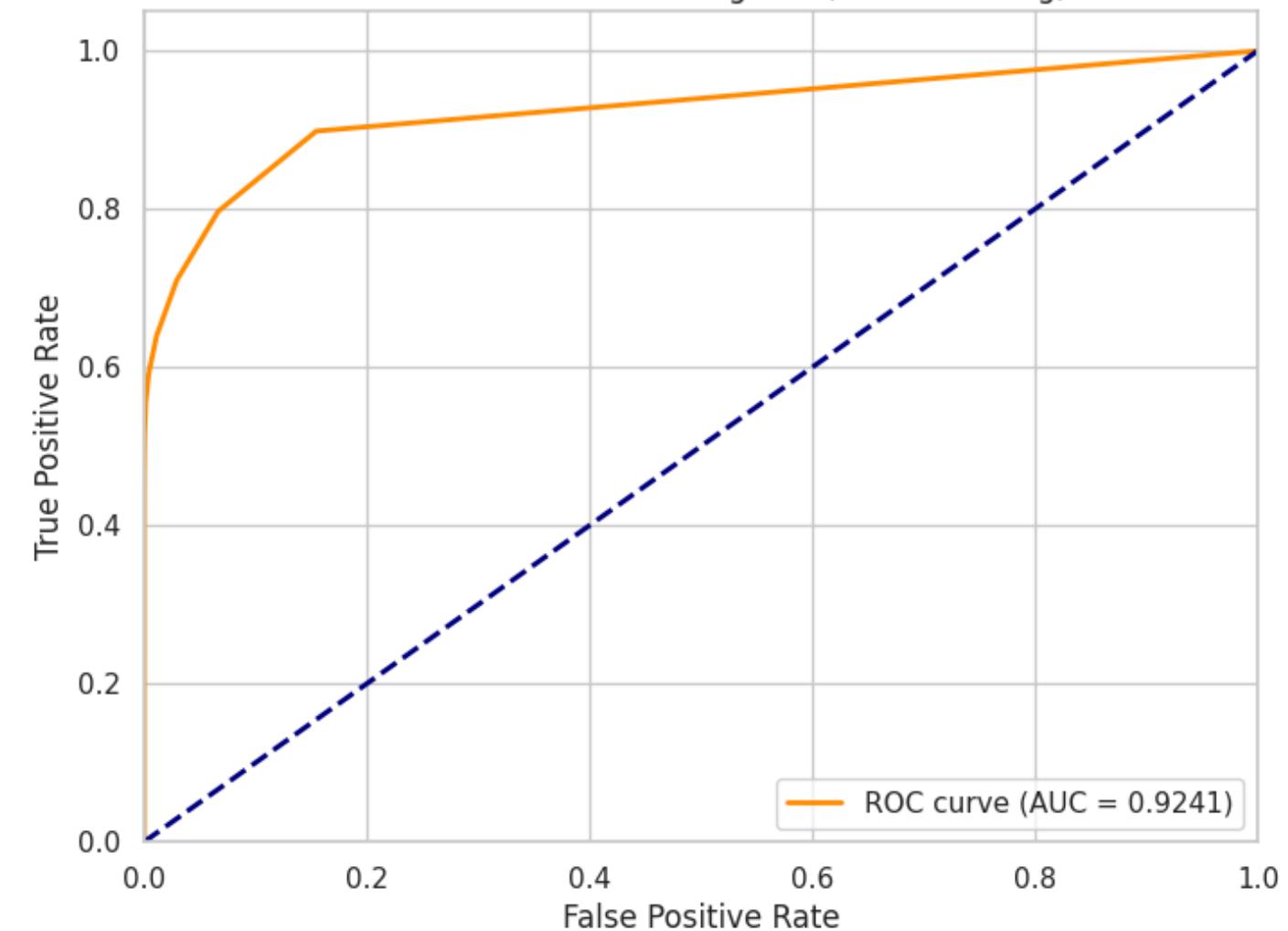
Recall: 0.5920

F1 Score: 0.7239

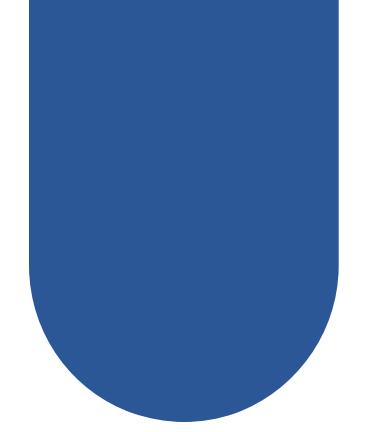
Confusion Matrix - K-Nearest Neighbor (Setelah Tuning)



ROC Curve - K-Nearest Neighbor (Setelah Tuning)



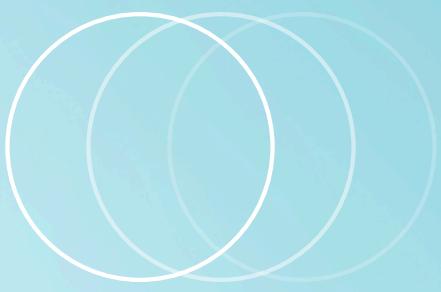
Tuning hyperparameter berhasil meningkatkan akurasi model K-Nearest Neighbor dari 0.9594 menjadi 0.9602. Model dengan hyperparameter optimal memiliki precision yang lebih tinggi tetapi recall yang sedikit lebih rendah dibandingkan model sebelumnya.



KESIMPULAN



Berdasarkan analisis dataset Diabetes Prediction, model terbaik adalah K-Nearest Neighbor dengan akurasi 96.02% setelah tuning hyperparameter. Fitur paling berpengaruh adalah HbA1c, glukosa darah, dan usia. Meskipun akurat, model menunjukkan trade-off antara precision (93.14%) dan recall (59.20%), sehingga lebih baik dalam meminimalkan false positive. Teknik preprocessing seperti penanganan outlier, encoding, dan normalisasi turut meningkatkan performa model.



THANK YOU

FOR YOUR ATTENTION

