Nama/NIM :Rafli Limandijaya/1103210243

1. Feature Engineering, EDA, dan Visualisasi

Feature Engineering

- Menghapus kolom yang tidak relevan atau terlalu banyak missing values.
- Menangani missing values: menghapus baris dengan two_year_recid kosong, atau drop kolom dengan missing value parah.
- Encoding data kategorik (misalnya: race, sex, c_charge_degree) menggunakan LabelEncoder atau OneHotEncoder.
- Menyusun kembali fitur numerik seperti priors_count, age, dll agar lebih representatif.

EDA & Visualisasi

- Visualisasi distribusi target (two_year_recid) untuk mengetahui ketidakseimbangan kelas.
- Korelasi antar fitur numerik menggunakan heatmap.
- Analisis pengaruh sex, race, priors_count, dan age terhadap recidivism menggunakan grafik batang dan histogram.
- Pendeteksian outlier & nilai ekstrem.

2. Evaluasi Model: Matriks dan Penjelasan

Berikut metrik evaluasi klasifikasi yang digunakan dan penjelasan matematisnya:

a. Accuracy

Akurasi mengukur proporsi prediksi yang benar terhadap total prediksi.

2. Evaluasi Model: Matriks dan Penjelasan

Berikut metrik evaluasi klasifikasi yang digunakan dan penjelasan matematisnya:

a. Accuracy

Akurasi mengukur proporsi prediksi yang benar terhadap total prediksi.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

b. Precision

Precision menunjukkan seberapa akurat model saat memprediksi kelas positif. Cocok untuk konteks ketika *False Positive mahal*.

$$\text{Precision} = \frac{TP}{TP + FP}$$

c. Recall mengukur seberapa banyak dari total kasus positif yang berhasil ditemukan oleh model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Berfungsi untuk meminimalkan false negative

d. F1 Score

Gabungan harmonis antara Precision dan Recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Digunakan ketika dataset tidak seimbang dan kita ingin trade-off yang adil antara precision dan recall.

e. AUC (Area Under the Curve)

AUC mengukur kemampuan model membedakan antara kelas positif dan negatif pada berbagai threshold.

Nilai AUC berada di antara 0 dan 1

f. ROC Curve (Receiver Operating Characteristic)

Grafik antara True Positive Rate (TPR) dan False Positive Rate (FPR).

$$\text{TPR} = \frac{TP}{TP + FN} \quad ; \quad \text{FPR} = \frac{FP}{FP + TN}$$

ROC digunakan untuk mengevaluasi kemampuan diskriminatif model di semua threshold.

3. Kesimpulan dari Evaluasi

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98       823
           1       0.95      1.00      0.97       620

    accuracy                           0.98      1443
   macro avg       0.98      0.98      0.98      1443
weighted avg       0.98      0.98      0.98      1443

Confusion Matrix:
 [[792  31]
 [  1 619]]
ROC AUC Score: 0.9940216752243954
```

- Akurasi tinggi bukan segalanya, terutama jika datanya tidak seimbang.
- Precision & Recall penting untuk memahami seberapa bagus model saat menghadapi kasus positif.
- AUC mendekati 1 artinya model sangat bagus membedakan dua kelas.
- F1 Score menyatukan presisi dan recall: ideal untuk dataset seperti COMPAS yang sensitif terhadap klasifikasi salah.