Evaluasi Model Klasifikasi dengan ROC dan AUC

Sebelumnya, kita telah membahas beberapa matrik evaluasi untuk model klasifikasi seperti akurasi, presisi, recall, dan F1-score. Hari ini, kita akan membahas dua matrik evaluasi lainnya yang sangat berguna, terutama ketika dataset tidak seimbang: Receiver Operating Characteristic (ROC) curve dan Area Under the Curve (AUC).

# ROC Curve

ROC curve adalah grafik yang menggambarkan kinerja model klasifikasi pada berbagai ambang batas klasifikasi. Kurva ini diplot dengan True Positive Rate (TPR) pada sumbu y dan False Positive Rate (FPR) pada sumbu x.

* **True Positive Rate (TPR):** Proporsi sampel positif yang diklasifikasikan dengan benar sebagai positif. Disebut juga recall atau sensitivity.
* **False Positive Rate (FPR):** Proporsi sampel negatif yang salah diklasifikasikan sebagai positif.

## Interpretasi ROC Curve

* Semakin dekat kurva ROC ke sudut kiri atas, semakin baik kinerja model.
* Kurva ROC yang berupa garis diagonal menunjukkan kinerja model yang acak (tidak lebih baik dari tebakan acak).

# AUC (Area Under the Curve)

AUC adalah luas area dibawah kurva ROC. AUC adalah metrik tunggal yang merangkum kinerja model klasifikasi pada semua ambang batas klasifikasi.

## Interpretasi AUC

* Nilai AUC berkisar antara 0 hingga 1.
* Semakin tinggi nilai AUC, semakin baik kinerja model dalam membedakan antara kelas positif dan negatif.
* AUC = 0.5 menunjukkan kinerja model yang acak.
* AUC = 1 menunjukkan kinerja model yang sempurna.

# Mengapa Menggunakan ROC dan AUC?

* **Dataset Tidak Seimbang:** ROC dan AUC sangat berguna ketika dataset tidak seimbang (jumlah sampel per kelas tidak sama). Metrik ini tidak dipengaruhi oleh distribusi kelas dalam dataset.
* **Membandingkan Model:** AUC adalah metrik yang baik untuk membandingkan kinerja beberapa model klasifikasi. Model dengan AUC yang lebih tinggi biasanya lebih baik.

**Tugas Hari 32**

Menggunakan dataset “Social Network Ads”

**Tugas:**

1. **Hitung Probabilitas:**

* Gunakan model regresi logistik yang telah di buat untuk menghitung probabilitas kelas positif (membeli produk) untuk setiap sampel data uji.

1. **Buat Kurva ROC:**

* Gunakan fungsi roc\_curve dari scikit-learn untuk membuat kurva ROC berdasarkan probabilitas kelas positif dan label sebenarnya (y\_test).

1. **Hitung AUC:**

* Gunakan fungsi auc dari scikit-learn untuk menghitung AUC.

1. **Visualisasi:**

* Plot Kurva ROC.
* Tampilkan nilai AUC pada plot

**Contoh Kode (Scikit-Learn dan Matplotlib)**

from sklearn.metrics import roc\_curve, auc

# ... (kode untuk membaca, mempersiapkan data, membagi data, dan membuat model)

# Menghitung probabilitas kelas positif

y\_prob = model.predict\_proba(X\_test)[:, 1]

# Menghitung ROC curve dan AUC

fpr, tpr, thresholds = roc\_curve(y\_test, y\_prob)

roc\_auc = auc(fpr, tpr)

# Plot ROC curve

plt.figure()

plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area = {roc\_auc:.2f})')

plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')

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')

plt.legend(loc="lower right")

plt.show()

**Selamat Mengerjakan Tugas! 🙂**

**Tugas**

1. **Hitung Probabilitas:**

**from sklearn.metrics import roc\_curve, auc**

**# Menghitung Probabilitas Kelas Positif**

**y\_prob = model.predict\_proba(X\_test)[:, 1]**

**print(y\_prob)**

1. **Buat Kurva ROC:**

**# Buat ROC**

**fpr, tpr, thresholds = roc\_curve(y\_test, y\_prob)**

**print("\nFPR:",fpr)**

**print("\nTPR:",tpr)**

**print("\nTHRESHOLDS:",thresholds)**

1. **Hitung AUC:**

**# Hitung AUC**

**roc\_auc = auc(fpr, tpr)**

**print("\nHasil AUC:",{roc\_auc})**

1. **Visualisasi:**

**# Visualisasi plot ROC curve**

**plt.figure()**

**plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area = {roc\_auc:.2f})')**

**plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')**

**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')**

**plt.legend(loc="lower right")**

**plt.show()**

**Penjelasan Kode:**

1. **Hitung Probabilitas:**
   * y\_prob = model.predict\_proba(X\_test)[:, 1]: Menggunakan metode predict\_proba dari model regresi logistik yang sudah dilatih untuk mendapatkan probabilitas setiap sampel data uji termasuk dalam kelas positif (membeli produk). [:, 1] digunakan untuk mengambil kolom kedua dari array hasil prediksi, yang merupakan probabilitas untuk kelas positif.
2. **Buat Kurva ROC:**
   * fpr, tpr, thresholds = roc\_curve(y\_test, y\_prob): Menghitung False Positive Rate (FPR), True Positive Rate (TPR), dan threshold yang sesuai untuk setiap titik pada kurva ROC.
   * print("\nFPR:", fpr): Mencetak nilai FPR.
   * print("\nTPR:", tpr): Mencetak nilai TPR.
   * print("\nTHRESHOLDS:", thresholds): Mencetak nilai threshold.
3. **Hitung AUC:**
   * roc\_auc = auc(fpr, tpr): Menghitung Area Under the Curve (AUC) dari kurva ROC.
4. **Visualisasi:**
   * plt.figure(): Membuat figure baru untuk plot.
   * plt.plot(fpr, tpr, ...): Membuat plot kurva ROC dengan warna orange dan label yang menunjukkan nilai AUC.
   * plt.plot([0, 1], [0, 1], ...): Membuat garis diagonal sebagai referensi kinerja model yang acak.
   * plt.xlim([0.0, 1.0]), plt.ylim([0.0, 1.05]): Mengatur batas sumbu x dan y.
   * plt.xlabel('False Positive Rate'), plt.ylabel('True Positive Rate'), plt.title('Receiver Operating Characteristic'): Memberikan label pada sumbu dan judul plot.
   * plt.legend(loc="lower right"): Menampilkan legenda.
   * plt.show(): Menampilkan plot.

**Interpretasi Hasil:**

* **FPR, TPR, Thresholds:**
  + FPR (False Positive Rate) adalah tingkat kesalahan di mana model memprediksi positif (membeli) padahal kenyataannya negatif (tidak membeli).
  + TPR (True Positive Rate) adalah tingkat keberhasilan model dalam memprediksi positif (membeli) dengan benar.
  + Thresholds adalah nilai ambang batas probabilitas yang digunakan untuk memutuskan apakah suatu sampel diklasifikasikan sebagai positif atau negatif.
* **AUC:**
  + Nilai AUC biasanya berkisar antara 0,5 (kinerja acak) hingga 1,0 (kinerja sempurna).
  + Semakin tinggi nilai AUC, semakin baik kemampuan model dalam membedakan antara kelas positif dan negatif.
* **Visualisasi ROC Curve:**
  + Kurva ROC menunjukkan hubungan antara TPR dan FPR pada berbagai threshold.
  + Kurva yang lebih tinggi dan mendekati sudut kiri atas menunjukkan kinerja model yang lebih baik.

Dengan menggunakan ROC dan AUC, Anda dapat mengevaluasi model klasifikasi secara lebih komprehensif, terutama ketika dealing dengan dataset yang tidak seimbang.