Assignment - Individual	
Course	**Pengantar Pembelajaran Mesin** (3 SKS)
Lecturer	Dr. Bambang Heru Iswanto
Due date	(see Epsilon)
Instruction	• Tulis nama anda dan NIM • Jawaban dan program dengan Jupyter Notebook/Google Colab • Upload jawaban ke Epsilon dalam format: ipynb dan .html

Session:

Evaluasi Model

1. ***Confusion Matrix***. Anda memiliki dataset untuk membangun model klasifikasi apakah suatu email tergolong "spam" atau "non-spam". Setelah melatih model klasifikasi, Anda mendapatkan *confusion matrix* berikut:

	Actual	
Predicted	Spam	Non-Spam
Spam	90	10
Non-Spam	20	180

   a. Berapa jumlah email yang diklasifikasikan dengan benar oleh model Anda?
   b. Hitunglah akurasi, presisi, recall, dan F-Measure;
   c. Jelaskan apa arti dari nilai-nilai yang Anda peroleh dari hasil (c) ?

   Note: perhitungan boleh manual atau menggunakan paket program yang tersedia;

2. **Cross-Validation**. Misalkan Anda memiliki data berukuran 1000 sampel dengan 10 fitur yang digunakan untuk klasifikasi citra toraks menjadi dua kelas: "cancer" dan "non-cancer". Untuk evaluasi performa model, Anda ingin menggunakan validasi silang (*cross validation*) 5-fold.
   a. Jelaskan bagaimana proses validasi silang 5-fold dan jelaskan pula apa keuntungan menggunakan validasi silang dalam evaluasi model?
   b. Apabila ternyata model memiliki akurasi rata-rata sebesar 90% namun pada satu fold tertentu memiliki performa sangat buruk dengan akurasi hanya 60%, apa yang dapat Anda simpulkan dan bagaimana menanganinya?

3. **Receiver Operating Characteristic (ROC)**. Anda telah melatih sebuah model klasifikasi untuk memprediksi apakah seorang pasien memiliki penyakit X berdasarkan sejumlah fitur medis. Setelah melatih model, performa model dievaluasi menggunakan data uji dari 200 pasien. Berikut adalah probabilitas positif yang diprediksi oleh model untuk 20 pasien pertama:

   Probabilitas Positif: 0.75, 0.65, 0.82, 0.57, 0.41, 0.36, 0.93, 0.79, 0.68, 0.61,
   0.53, 0.49, 0.81, 0.72, 0.67, 0.74, 0.55, 0.48, 0.39, 0.71

   Anda memiliki label sebenarnya untuk setiap pasien, di mana Berikut adalah label sebenarnya untuk 20 pasien pertama, dimana "1" untuk pasien memiliki penyakit X dan "0" sebaliknya.

Label Sebenarnya: 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1

Jawablah pertanyaan-pertanyaan berikut:

a. Urutkan probabilitas positif dan label sebenarnya berdasarkan probabilitas positif yang diprediksi dari yang tertinggi hingga terendah.
b. Gambarkan kurva ROC menggunakan data ini.
c. Hitunglah nilai AUC (Area Under the Curve) dari kurva ROC.
d. Jika Anda harus memilih threshold probabilitas positif tertentu untuk mengklasifikasikan pasien sebagai positif atau negatif, bagaimana Anda akan memilih threshold tersebut berdasarkan kurva ROC?
