

a. Jelaskan apa yang dimaksud dengan hold-out validation dan k-fold cross-validation!

⇒ **Hold-out validation**: teknik untuk membagi dataset menjadi dua bagian utama: *training* set dan *testing* set. *Training* dataset digunakan untuk melatih model, sedangkan *testing* dataset digunakan untuk mengevaluasi performa model yang telah dilatih.

⇒ **Cross validation**: teknik untuk membagi dataset menjadi beberapa *fold* (lipatan) yang setiap *fold* akan digunakan secara bergantian sebagai *testing* set dan sisanya sebagai *training* set.

b. Jelaskan kondisi yang membuat hold-out validation lebih baik dibandingkan dengan k-fold cross-validation, dan jelaskan pula kasus sebaliknya!

	<i>Hold-out validation</i>	<i>K-fold cross-validation</i>
Dataset sangat besar	✓	
[Sebaliknya] Dataset relatif kecil		✓
Waktu yang diperlukan sangat terbatas	✓	
[Sebaliknya] Memiliki banyak waktu		✓
Fase awal pengembangan (untuk mendapatkan estimasi)	✓	
[Sebaliknya] Fase pengembangan tengah-akhir (untuk mendapatkan hasil yang lebih akurat)		✓
Menghindari <i>overfitting</i>		✓
Mendapatkan estimasi yang stabil		✓

Kemudahan implementasi	✓	
Efisiensi komputasi	✓	

c. Apa yang dimaksud dengan *data leakage*?

⇒ *Data leakage* merupakan fenomena dimana informasi dari data *testing* secara tidak sengaja “bocor” ke dalam proses pelatihan model sehingga akan menyebabkan model menjadi *overfit* dan menurunkan kualitas hasil yang dihasilkan.

d. Bagaimana dampak *data leakage* terhadap kinerja dari model?

⇒ Model menjadi *overfit*

⇒ Kualitas pengambilan keputusan model menjadi menurun

⇒ Akurasi yang dihasilkan tidak realistis (implikasi dari *overfit*)

e. Berikanlah solusi untuk mengatasi permasalahan *data leakage*!

⇒ Memisahkan data dengan baik dan jelas dengan melakukan pemisahan data menjadi data *training* dan *testing*.

⇒ Melakukan urutan proses secara runut dan terorganisir supaya bilamana ada data yang ditransformasi tidak akan bocor

⇒ Hindari penggunaan informasi dari masa depan, dalam artian jangan menggunakan informasi yang baru diketahui setelah terjadinya suatu peristiwa

⇒ Menggunakan *k-fold cross-validation* supaya dapat mengetahui hasilnya apakah stabil