Modeling

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

   Pada hold-out validation, data latih dan test dibagi lansung, biasanya dengan dibagi sebesar 80-20. Model dilatih menggunakan data latih dan kemudian performanya diukur dengan memprediksi data test.

   K-fold cross-validation bekerja mirip dengan hold-out, tapi dilakukan k kali. Pertama, tentukan nilai k. Kemudian, data dibagi secara acak menjadi k bagian yang ukurannya sama, bagian ini disebut dengan nama fold. Proses validasi seperti pada hold-out dilakukan secara berulang sebanyak k kali dengan dalam setiap putaran 1 fold menjadi data test dan k-1 folds lainnya menjadi data latih. Skornya dicatat untuk setiap putaran. Skor akhir didapatkan dari rata2 semua skor yang sudah dicatat.

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

   Hold-out validation lebih baik dilakukan jika dataset yang dimiliki sangat besar dan prosesnya tidak akan lama. Sebaliknya, k-fold cross-validation lebih baik digunakan ketika dataset kecil atau tidak terlalu besar karena dapat menghasilkan estimasi performa yang lebih handal.

3. **Apa yang dimaksud dengan *data leakage*?**

   Data leakage adalah ketika ada informasi dari data test yang bocor ke data latih. Hal ini biasanya terjadi ketika melakukan preprocessing sebelum membagi data, seperti scaling, dan terdapat duplikat pada data latih dan test.

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

   Evaluasi model tidak akurat karena model menjadi bias terhadap dataset latihan.

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

   Lakukan data cleaning untuk yang duplikat dan lakukan beberapa preprocessing data setelah pembagian data.