Seleksi Asisten Lab AI '23

Supervised Learning: Modeling Theory

---

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

   - **Hold-out Validation** adalah strategi validasi yang paling sederhana. Dataset dibagi menjadi dua bagian yang terpisah yaitu (*training set*) dan (*testing set*). Model dilatih hanya menggunakan *training set*, dan kinerjanya dievaluasi satu kali pada *testing set* yang belum pernah dilihat sebelumnya.
   - **K-Fold Cross-Validation** adalah strategi yang lebih robust. Dataset dibagi menjadi k bagian yang berukuran sama. Proses validasi kemudian diulang sebanyak k kali. Pada setiap iterasi, satu lipatan digunakan sebagai data pengujian, dan k-1 lipatan sisanya digunakan sebagai data pelatihan. Hasil kinerja dari k iterasi tersebut kemudian dirata-ratakan untuk mendapatkan estimasi kinerja model yang lebih stabil.

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

   - **Kondisi Hold-out Lebih Baik:** *Hold-out validation* akan menjadi lebih baik ketika kita memiliki dataset yang sangat besar. Pada kasus ini, membuat satu *testing set* yang representatif saja sudah cukup. Menjalankan *k-fold cross-validation* pada dataset yang besar akan memakan waktu dan biaya komputasi yang sangat mahal tanpa memberikan manfaat tambahan yang signifikan.
   - **Kondisi K-Fold Lebih Baik:** *K-fold cross-validation* akan menjadi lebih baik ketika kita bekerja dengan dataset yang ukurannya kecil atau sedang. Karena pada dataset kecil, pembagian acak pada metode *hold-out* bisa sangat berpengaruh pada hasil (misalnya, secara kebetulan data yang sulit terkumpul di *testing set*). *K-fold* mengatasi ini dengan memastikan setiap titik data pernah menjadi bagian dari *training set* dan *testing set*, sehingga memberikan evaluasi kinerja yang jauh lebih dapat diandalkan dan tidak terlalu bias oleh pembagian data.

3. Apa yang dimaksud dengan *data leakage*?

Data leakage atau kebocoran data adalah kesalahan dalam proses perancangan model dimana informasi dari luar data pelatihan "bocor" ke dalam proses pelatihan model. Artinya model secara tidak sengaja "mengintip" data pengujiannya. Hal ini menciptakan persepsi yang salah, di mana model tampak memiliki kinerja yang sangat tinggi saat evaluasi karena sudah mengintip, tetapi gagal ketika dihadapkan pada data baru.

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

Dampak utama dari *data leakage* adalah evaluasi kinerja yang terlalu bagus dan tidak realistis. Model akan menunjukkan skor akurasi, presisi, atau metrik lainnya yang sangat tinggi, padahal kinerja ini palsu. Ketika model tersebut digunakan di data yang baru, kinerjanya akan turun drastis. Hal ini menyebabkan model menjadi tidak dapat diandalkan dan gagal mencapai tujuan bisnisnya.

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

Beberapa solusi fundamental untuk mencegah *data leakage* adalah:

- **Pemisahan Data yang Jelas:** Lakukan pemisahan data menjadi *training set* dan *testing set* di awal proses, sebelum melakukan pemrosesan apa pun. Jangan pernah menyentuh *testing set* sampai tahap evaluasi akhir.
- **Menggunakan Pipeline:** Kita bisa memanfaatkan fitur *Pipeline* dari *library* seperti Scikit-learn untuk merangkai langkah pra-pemrosesan dan pelatihan model. *Pipeline* secara otomatis memastikan bahwa setiap langkah dijalankan dengan benar dan mencegah kebocoran data, terutama saat menggunakan *cross-validation*.