SOAL

1. Jelaskan apa yang dimaksud dengan *hold-out validation* dan *k-fold cross-validation*!
2. Jelaskan kondisi yang membuat *hold-out validation* lebih baik dibandingkan dengan *k-fold cross-validation*, dan jelaskan pula kasus sebaliknya!
3. Apa yang dimaksud dengan *data leakage*?
4. Bagaimana dampak *data leakage* terhadap kinerja dari model?
5. Berikanlah solusi untuk mengatasi permasalahan *data leakage*!

JAWABAN

1. **Hold-Out Validation**
   Hold-out validation adalah metode validasi di mana dataset dibagi menjadi dua subset. Satu subset digunakan untuk melatih model (training set), dan subset lainnya digunakan untuk menguji model (test set). Biasanya, dataset dibagi dengan rasio tertentu, misalnya 80% untuk training dan 20% untuk testing.
   Setelah model dilatih menggunakan training set, performanya diukur menggunakan test set untuk melihat bagaimana model dapat di-*generalize* pada data yang belum pernah dilihat sebelumnya.
   K-Fold Cross-Validation
   K-fold cross-validation adalah metode validasi yang membagi dataset menjadi K bagian (folds) yang sama besar. Model dilatih K kali, setiap kali menggunakan K-1 bagian sebagai training set dan 1 bagian sebagai test set. Hasil dari K model ini kemudian dirata-rata untuk menghasilkan estimasi kinerja model yang lebih stabil dan mengurangi variabilitas akibat pembagian dataset.

2. **Hold-Out Validation Lebih Baik**
   a. **Waktu dan Sumber Daya Terbatas**
      Jika komputasi dan waktu adalah faktor penting, hold-out validation lebih efisien karena model hanya perlu dilatih dan diuji sekali. Ini bisa sangat penting dalam skenario real-time atau ketika modelnya sangat kompleks.
   b. **Dataset Sangat Besar**
      Ketika dataset sangat besar, hold-out validation dapat memberikan estimasi yang baik terhadap kinerja model karena subset yang digunakan untuk testing masih cukup representatif dari keseluruhan dataset.

   K-Fold Cross-Validation Lebih Baik

a. **Dataset Kecil**: K-fold cross-validation lebih efektif saat bekerja dengan dataset kecil, karena metode ini memungkinkan setiap bagian dari data digunakan untuk pengujian, memberikan estimasi yang lebih stabil dan mengurangi risiko pembagian data yang tidak representatif.
b. **Evaluasi Model yang Lebih Akurat**: Jika tujuan adalah untuk mendapatkan evaluasi yang lebih akurat dari performa model, k-fold cross-validation lebih diinginkan karena hasilnya didasarkan pada beberapa eksperimen, mengurangi variabilitas hasil yang disebabkan oleh pembagian dataset secara acak.

3. Data leakage terjadi ketika informasi diluar dari training dataset digunakan untuk membuat model sehingga model memiliki akses ke informasi yang seharusnya tidak tersedia saat model diterapkan pada data baru. Ini biasanya terjadi karena data test atau target secara tidak sengaja digunakan dalam proses pelatihan, menyebabkan model untuk "curang" dan menghasilkan kinerja yang tampak lebih baik dari yang seharusnya.

4. Data leakage dapat memberikan kinerja model yang terlalu optimistis selama pelatihan dan validasi karena model telah melihat atau menggunakan informasi dari set yang seharusnya tidak diketahui selama pelatihan. Ini akan menyebabkan model untuk menunjukkan hasil yang jauh lebih baik saat diuji pada data yang telah bocor, tetapi akan memiliki performa yang buruk pada data baru atau *unseen* data. Ini mengarah pada *overfitting*, suatu kasus pada model yang tampaknya memiliki performa yang sangat baik pada data pelatihan dan validasi, tetapi gagal dalam generalisasi ke data baru yang belum pernah dilihatnya.

5. Berikut adalah Langkah untuk mencegah *data leakage*:

    a. **Pisahkan Data dengan Benar**
    Pastikan bahwa data test atau data yang seharusnya tidak digunakan dalam pelatihan benar-benar dipisahkan dari data pelatihan.

    b. **Konsisten dengan Data Preprocessing**
    Gunakan parameter preprocessing (seperti mean atau standard deviation dalam scaling) yang diperoleh dari data training dan terapkan secara konsisten pada data test tanpa melakukan *fitting* kembali pada data test.