Rizqika Mulia Pratama
13522126

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

   Hold-out validation dan k-fold cross-validation adalah dua teknik yang sering digunakan dalam evaluasi model machine learning, yang bertujuan untuk menilai seberapa baik model tersebut dapat digeneralisasi ke data yang belum pernah dilihatnya. Dalam hold-out validation, dataset yang dibagi menjadi dua bagian terpisah: satu bagian digunakan untuk melatih model, yang disebut dengan training set, dan bagian lainnya digunakan untuk menguji model, yang disebut test set. Pembagian ini biasanya dilakukan dengan proporsi seperti 70-30 atau 80-20, di mana sebagian besar data digunakan untuk pelatihan. Metode ini sederhana dan cepat untuk diterapkan, tetapi memiliki kelebihan karena hasil evaluasi sangat bergantung pada bagaimana data dibagi. Jika pembagian data tidak representatif, hasil evaluasi bisa menjadi bias dan tidak akurat.

   Sebaliknya, k-fold cross-validation menawarkan pendekatan yang lebih komprehensif dan akurat. Dalam metode ini, dataset dibagi menjadi k bagian atau “folds” yang kurang lebih sama besar. Model kemudian dilatih dan diuji sebanyak k kali, di mana setiap kali satu fold digunakan sebagai test set dan k-1 fold lainnya digunakan sekali sebagai test set, dan hasil evaluasi model adalah rata-rata dari k percobaan tersebut. Kelebihan dari k-fold cross-validation adalah kemampuannya untuk memberikan penilaian yang lebih stabil dan akurat, karena mengurangi risiko bias yang disebabkan oleh pembagian data yang tidak representatif. Namun, metode ini lebih memakan waktu dan sumber daya komputasi karena model harus dilatih dan diuji berkali-kali. Kedua teknik ini sangat penting dalam memastikan bahwa model machine learning yang dibangun memiliki kinerja yang baik dan dapat diandalkan ketika diterapkan pada data baru.

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

   Hold-out validation dan k-fold cross-validation masing-masing memiliki kelebihan dan kekurangan yang membuatnya lebih cocok untuk situasi tertentu.

   Hold-out validation lebih baik ketika:

   1. Dataset sangat besar

      Jika dataset sangat besar, hold-out validation bisa lebih efisien. Dengan data yang melimpah, pembagian sederhana menjadi training dan test set sudah cukup untuk memberikan evaluasi yang representatif. Dalam kasus ini, kelebihan hold-out validation adalah kecepatan dan kemudahan implementasinya, karena hanya memerlukan satu kali pelatihan dan pengujian.

2. Keterbatasan waktu dan sumber daya

   Jika waktu dan sumber daya komputasi terbatas, hold-out validation bisa menjadi pilihan yang lebih praktis. Karena hanya melibatkan satu kali pelatihan dan pengujian, metode ini jauh lebih cepat dan tidak memerlukan daya komputasi yang besar dibandingkan dengan k-fold cross-validation.

K-fold Cross-Validation lebih baik ketika:

1. Dataset kecil atau terbatas

   Ketika dataset berukuran kecil, k-fold cross-validation lebih disukai karena memaksimalkan penggunaan data. Dengan membagi data menjadi beberapa fold dan melakukan pelatihan serta pengujian sebanyak k kali, metode ini memastikan bahwa setiap data point digunakan untuk pelatihan dan pengujian, sehingga memberikan estimasi kinerja model yang lebih akurat.

2. Menghindari bias pembagian data

   Dalam situasi di mana ada risiko bahwa pembagian data secara acak dapat menghasilkan subset yang tidak representatif, k-fold cross-validation membantu mengurangi bias ini. Dengan melakukan evaluasi k kali dan menggunakan rata-rata hasilnya, metode ini memberikan penilaian yang lebih stabil dan mengurangi variabilitas yang disebabkan oleh pembagian data yang tidak rata.

Secara umum, pilihan hold-out validation dan k-fold cross-validation bergantung pada ukuran dataset, ketersediaan sumber daya, dan kebutuhan akan akurasi evaluasi. Hold-out validation lebih cocok untuk situasi di mana kecepatan dan efisiensi menjadi prioritas, sementara k-fold cross-validation lebih tepat digunakan ketika akurasi dan keandalan evaluasi menjadi fokus utama, terutama pada dataset yang lebih kecil.

3. Apa yang dimaksud dengan *data leakage*?

Data leakage adalah masalah serius dalam machine learning yang terjadi ketika informasi dari luar dataset pelatihan secara tidak sengaja digunakan untuk membangun model, sehingga memberikan model akses ke data yang seharusnya tidak tersedia selama pelatihan. Ini dapat menyebabkan model

tampak memiliki kinerja yang sangat baik selama fase pelatihan dan validasi, tetapi kinerjanya menurun drastis ketika diterapkan pada data baru yang belum pernah dilihat sebelumnya

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

   Data leakage dapat secara signifikan merusak kinerja model machine learning dengan memberikan ilusi kinerja yang lebih baik selama pelatihan dan validasi. Ketika informasi yang seharusnya tidak tersedia untuk model bocor ke dalam proses pelatihan, model dapat "mempelajari" pola yang tidak akan ada dalam data baru, sehingga metrik evaluasi seperti akurasi tampak lebih tinggi dari sebenarnya. Akibatnya, ketika model diterapkan pada data baru, kinerjanya seringkali menurun drastis karena tidak dapat digeneralisasi dengan baik. Hal ini dapat menyebabkan keputusan bisnis yang salah, seperti dalam prediksi risiko kredit, di mana model tampaknya akurat dapat memberikan hasil yang menyesatkan.

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

   Pertama, pastikan pembagian data dilakukan dengan benar. Pisahkan dataset menjadi training, validation, dan test set sebelum melakukan preprocessing apapun. Ini mencegah informasi dari test set bocor ke dalam training set. Kedua, lakukan preprocessing seacar terpisah untuk setiap subset data. Misalnya, normalisasi atau pengisian nilai yang hilang harus dilakukan hanya pada training set, dan kemudian menerapkan transformasi yang sama pada validation dan test set. Ketiga, berhati-hati dalam pemilihan fitur. Pastikan bahwa fitur yang digunakan tidak mengandung informasi target atau informasi yang hanya tersedia di masa depan. Terakhir, gunakan teknik validasi yang tepat, seperti k-fold cross-validation, untuk memastikan bahwa model dievaluasi secara menyeluruh dan tidak ada informasi yang bocor antar fold.