Julian Chandra Sutadi - 13522080

Pertanyaan

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 adalah strategi pemisahan dataset menjadi dua: train dan test. Salah satu proporsi pemisahan yang umum adalah 80% dataset untuk train dan 20% untuk test. Model dilatih atas dataset train, kemudian error dihitung berdasarkan dataset train untuk mengevaluasi keberhasilan model.
   K-fold cross-validation adalah strategi untuk memisah dataset menjadi k subset, dengan salah satu subset sebagai test set. Model dilatih sebanyak k-kali dan error yang dihitung adalah berdasarkan rata-rata error dari tiap pelatihan model.
2. Hold-out validation dapat lebih baik jika dataset sangat besar, sehingga melakukan iterasi sebanyak k-kali akan memerlukan compute yang terlalu besar dan memakan sumber daya berlebihan. Namun, k-fold cross-validation baik untuk melakukan proses fitting dan mendapatkan parameter model final yang ideal, sebab jika dilakukan dalam proses cross-validation, dapat membantu mengurangi overfitting.
3. Data leakage adalah keadaan ketika data yang digunakan di dalam proses melatih model mengandung informasi tentang target, namun data yang serupa tidak akan tersedia ketika model digunakan untuk prediksi.
   Salah satu bentuk data leakage adalah target leakage, yaitu ketika suatu pengetahuan a posteriori terdapat dalam train data. Bentuk lainnya adalah kontaminasi train-test, ketika informasi dari dataset test terdapat di dalam dataset train, karena melakukan preprocessing sebelum pemisahan train-test. Salah satu contoh kontaminasi ini adalah melakukan normalisasi sebelum train-test split.
4. Data leakage akan menyebabkan overfitting, karena model dilatih menggunakan data yang mengandung informasi yang “terlalu membantu,” informasi yang tidak tersedia pada saat melakukan prediksi.
5. Untuk mengatasi data leakage akibat target leakage, data scientist dapat menghitung akurasi cross validation dari suatu model. Jika nilainya terlalu besar dan mencurigakan, maka dapat dicari variabel apa yang korelasinya sangat besar dengan target, dan variable tersebut dapat di drop. Untuk mengatasi kontaminasi train-test, pemisahan dataset harus dilakukan sejak awal.