Francesco Michael Kusuma

13522038

1. **Jelaskan apa yang dimaksud dengan *hold-out validation* dan *k-fold cross-validation*!**
   Hold-Out Validation adalah metode yang membagi dataset menjadi dua subset, yaitu *training set dan test set.* Lalu, model dilatih dengan *training set* dan dievaluasi dengan *test set.* Pembagian ini biasanya dilakukan secara acak, misalnya 80% untuk *training set* dan 20% untuk *test set*. Metode ini mudah untuk diimplementasikan dan cepat dieksekusi. Namun, hasil evaluasi bisa sangat bergantung pada cara dataset dibagi.

   K-Fold Cross-Validation adalah metode yang membagi dataset menjadi k subset yang sama besar. Lalu, model dilatih dan diuji k kali. Pelatihan ini dilakukan dengan menggunakan k-1 subset untuk *training* dan 1 subset tersisa untuk *testing*. Kemudian, hasil evaluasi diambil dengan menghitung rata-rata setelah k iterasi. Metode ini cenderung memberi performa yang lebih akurat karena memanfaatkan seluruh data untuk training dan testing.

2. **Jelaskan kondisi yang membuat *hold-out validation* lebih baik dibandingkan dengan *k-fold cross-validation*, dan jelaskan pula kasus sebaliknya!**
   Hold-Out Validation:
   a. Dataset yang dimiliki sangat besar dan pembagian sederhana sudah cukup merepresentasikan model seluruh populasi.
   b. Memerlukan kesepatan pemrosesan karena *hold-out validation* hanya memerlukan sekali pembagian data dan pelatihan model.
   c. Melakukan eksperimen awal untuk mendapatkan gambaran kasar tentang kinerja model sebelum menggunakan metode evaluasi yang lebih baik.

   K-Fold Cross-Validation:
   a. Dataset yang dimiliki kecil dan terbatas sehingga dapat memaksimalkan penggunaan data yang tersedia untuk pelatihan dan evaluasi.
   b. Perlu mengurangi variabilitas hasil yang bergantung pada satu pembagian data.
   c. Dalam kasus model yang cenderung *overfitting*, *k-fold cross-validation* lebih berguna karena memberikan gambaran yang lebih lengkap tentang performa model di berbagai subset data.

3. **Apa yang dimaksud dengan *data leakage*?**
   Data leakage adalah situasi saat data yang berasal dari luar *training set* ikut masuk ke dalam proses pelatiahan. Ini sering terjadi ketika data yang seharusnya hanya ada dalam *testing* secara tidak sengaja digunakan dalam proses pelatihan model.

4. **Bagaimana dampak *data leakage* terhadap kinerja dari model?**
   a. *Overfitting*
      Model mungkin akan menunjukkan performa yang sangat baik pada training set dan testing set, tetapi gagal saat dihadapkan pada data baru karena telah mempelajari informasi yang tidak seharusnya diketahui.
   b. Evaluasi yang tidak akurat
      Hasil evaluasi yang dihasilkan bisa memiliki bias yang sangat tinggi dan memberi kesan model lebih baik dari yang sebenarnya.

c. Kesalahan pengambilan keputusan
   Adanya *overfitting* dan evaluasi model yang tidak akurat dapat membuat terjadinya kesalahan dalam pengambilan keputusan, baik untuk keperluan bisnis ataupun medis.

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

a. Pisahkan data dengan jelas
   Pisahkan secara jelas data untuk pelatihan, validasi, dan pengujian.
b. Lakukan *feature engineering* dengan tepat
   Ketika melakukan *feature engineering*, pastikan hanya menggunakan data yang tersedia pada saat prediksi dilakukan.
c. *Cross-Validation* dengan benar
   Pastikan tidak ada informasi yang bocor antar subset data yang digunakan untuk training dan testing.