Modeling

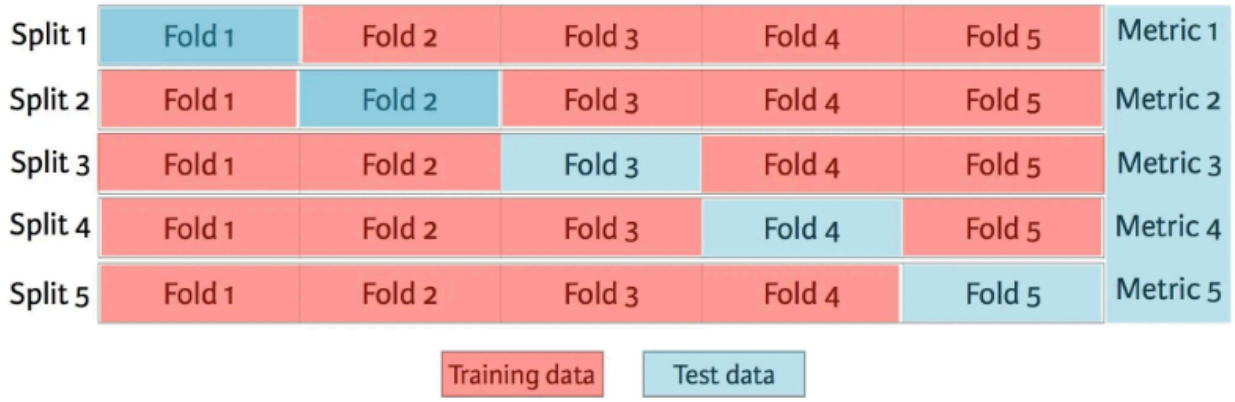
Adinda Putri - 13523071

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

Jawaban:

Hold-out validation merupakan pembagian data menjadi training dan test set. Training set digunakan untuk melatih model, sedangkan test set digunakan untuk melihat seberapa baik model bekerja pada data baru (unseen data). Umumnya, pembagian hold-out validation ini menggunakan rasio 80% untuk data training dan 20% untuk data testing.

K-Fold Cross-Validation merupakan pembagian data menjadi *k* bagian (fold) secara acak. Satu bagian digunakan untuk test set sementara sisanya digunakan untuk training set. Skor model terhadap test set tersebut dihitung. Lalu, proses ini dilakukan terus hingga setiap bagian unik telah digunakan sebagai test set. Ilustrasinya adalah sebagai berikut.



Gambar 1. K-Fold Cross-Validation

Sumber:https://medium.com/@jaz1/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f

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

Jawaban:

Hold-out validation lebih baik ketika:

- Initial exploratory analysis yang membutuhkan simplicity. Pendekatan ini *straightforward* dan mudah dipahami.
- Model yang digunakan mahal secara komputasi (misalnya deep learning atau dataset besar), karena hanya melakukan training sekali
- Dataset besar. Ketika data sangat banyak, performa model tidak terlalu sensitif terhadap subset yang dipilih sehingga hold-out validation cukup representatif

K-Fold Cross-Validation lebih baik ketika:

- Dataset kecil. Semua data digunakan untuk training dan test secara bergantian sehingga memanfaatkan keterbatasan data.
- Dataset balance. Ketika dataset balance, tiap fold mungkin untuk mewakili distribusi kelas keseluruhan.
- Hyperparameter tuning. k-Fold cocok digunakan untuk memilih hyperparameter karena memberikan estimasi performa model yang lebih stabil

c. Apa yang dimaksud dengan data leakage?

Jawaban:

Data leakage merupakan kebocoran data yang terjadi ketika model tidak sengaja menggunakan informasi pada test set saat training, sehingga mempengaruhi performa dan hasil evaluasi model. Akibatnya, model bekerja lebih baik namun tidak realistis kareka model dites untuk data yang pernah ia “lihat”. Data leakage juga dapat terjadi ketika model dilatih dengan informasi pada data training yang tidak tersedia saat test atau prediction. Hal ini bisa jadi sumber leakage konseptual karena model belajar hal yang tidak berlaku di dunia nyata.

d. Bagaimana dampak data leakage terhadap kinerja model?

Jawaban:

Beberapa dampak dari data leakage terhadap kinerja model adalah sebagai berikut.

- Generalisasi yang buruk pada data baru: Ketika model dilatih menggunakan informasi yang tidak mewakili dunia nyata, model akan kesulitan dalam menggeneralisasi data yang belum pernah dilihat.
- Biased decision-making: Bias pada data leakage menghasilkan decision yang unfair dan tidak realistis terhadap real-world scenarios.
- Inflated performance metrics: Data leakage membuat model menunjukkan kinerja baik, misalnya accuracy dan precision tinggi, yang "palsu".

e. Berikan solusi untuk mengatasi permasalahan data leakage!
Jawaban:

- Hindari mengekstrak fitur yang mengandung informasi mengenai target, pastikan fitur-fitur tersebut merefleksikan hanya yang tersedia saat prediction.
- Lakukan data preprocessing untuk training set dan test set secara terpisah.
- Lakukan data splitting secara tepat, tambahkan validation set diperlukan.
- Untuk time-series data, hindari penggunaan future untuk melatih model. Dapat digunakan teknik seperti rolling window validation atau walk-forward validation.
- Gunakan K-Fold Cross-Validation untuk mengidentifikasi potensi data leakage.

Referensi:

[1] *From Hold-Out to k-Fold: Understanding Cross-Validation Methods in Machine Learning*, Medium oleh Kavyasri Relangi, 2022. [Daring]. Tersedia: https://medium.com/@kavyasrirelangi100/from-hold-out-to-k-fold-understanding-cross-validation-methods-in-machine-learning-37402f406759. [Diakses: 5 September 2025].

[2] *What Is Data Leakage in Machine Learning?*, IBM Think, 30 September 2024. [Daring]. Tersedia: https://www.ibm.com/think/topics/data-leakage-machine-learning. [Diakses: 5 September 2025].

[3] *Data Leakage and Its Effect on the Performance of an ML Model*, Analytics Vidhya, 5 Juli 2021. [Daring]. Tersedia: https://www.analyticsvidhya.com/blog/2021/07/data-leakage-and-its-effect-on-the-performance-of-an-ml-model/. [Diakses: 5 September 2025].