Untuk mencapai balanced accuracy yang maksimal

1. Preprocessing data

* Pastikan data bersih dari nilai yang hilang (missing values), duplikat, atau anomali.
* Normalisasikan atau standardisasikan fitur-fitur jika diperlukan, terutama jika model yang digunakan sensitif terhadap skala data (misalnya, SVM atau k-NN).

1. Feature engineering

* Ciptakan fitur baru yang bisa memberikan informasi tambahan. Gunakan teknik seperti PCA atau Lasso Regression untuk memilih fitur yang paling relevan, guna menghindari overfitting.

1. Penanganan ketidakseimbangan kelas

* Gunakan teknik oversampling (misalnya SMOTE) atau undersampling untuk menyeimbangkan distribusi kelas.
* Saat melatih model, beri bobot lebih pada kelas minoritas untuk mengurangi bias model terhadap kelas mayoritas.
* Cara memberi bobot pada kelas minoritas:

1. Logistic regression
2. Random forest
3. Support vector machine (SVM)
4. Manual class weights
5. Pemilihan model

* Pilih model yang cocok untuk data dan masalah yang dihadapi. Coba beberapa model seperti Random Forest, Gradient Boosting, SVM, atau Neural Networks.
* Gunakan teknik ensemble seperti bagging atau boosting untuk meningkatkan kinerja prediksi.

1. Penyesuaian hyperparameter

* Lakukan pencarian grid atau random search untuk menemukan kombinasi hyperparameter yang optimal.
* Gunakan k-fold cross-validation untuk memastikan model tidak overfitting dan generalisasi dengan baik ke data yang belum pernah dilihat

1. Validasi dan Evaluasi

* Analisis confusion matrix untuk memahami performa model secara rinci
* Fokus pada metrik yang relevan seperti True Positive Rate (Sensitivity) dan True Negative Rate (Specificity)
  + True Positive (TP): Jumlah sampel positif yang berhasil diklasifikasikan dengan benar sebagai positif oleh model.
  + False Positive (FP): Jumlah sampel negatif yang secara keliru diklasifikasikan sebagai positif oleh model.
  + True Negative (TN): Jumlah sampel negatif yang berhasil diklasifikasikan dengan benar sebagai negatif oleh model.
  + False Negative (FN): Jumlah sampel positif yang secara keliru diklasifikasikan sebagai negatif oleh model.

1. Iterasi dan penyempurnaan

* Terus iterasi model dengan feedback dari evaluasi. Setiap kali lakukan perbaikan kecil dan evaluasi ulang.
* Lakukan iterasi pada pemilihan fitur jika performa tidak memuaskan.

1. Submission preparation

* Pastikan pipeline dari preprocessing, feature engineering, modeling, hingga evaluasi adalah stabil dan dapat direproduksi
* Simpan sebagian data sebagai data uji (holdout) yang benar-benar tidak terlihat oleh model untuk evaluasi akhir sebelum pengiriman

**GLOSARIUM**

**Oversampling**

Teknik meningkatkan jumlah sampel dalam kelas minoritas sehingga menjadi seimbang dengan kelas mayoritas

**Undersampling**

Undersampling adalah teknik untuk mengurangi jumlah sampel dalam kelas mayoritas sehingga menjadi seimbang dengan kelas minoritas

**Ensemble Learning**

Ensemble learning adalah teknik dalam pembelajaran mesin di mana beberapa model (disebut sebagai "pembelajar" atau "konstituen") digabungkan untuk menghasilkan hasil prediksi yang lebih baik daripada hasil yang dapat dicapai oleh model individual. Pendekatan ini memanfaatkan kekuatan kolektif dari beberapa model untuk meningkatkan akurasi, stabilitas, dan generalisasi dari sistem pembelajaran mesin.

Jenis-jenis ensemble learning:

1. Bagging (Bootstrap Aggregating)
2. Boosting
3. Stacking (Stacked Generalization)
4. Voting