Panduan Praktikum Classification (Ensemble Method & Class Imbalance)

Durasi: 3 × 110 menit

Capaian Pembelajaran Mata Kuliah

Mahasiswa mampu menjelaskan teknik penanganan class imbalance problem dan metode klasifikasi ensemble serta penggunaannya pada permasalahan klasifikasi.

Tools

- Google Colab
- Jupyter Notebook
- PyCharm
- Spyder
- Python IDE yang lain

Materi Praktikum

- Pengertian Classification

Classification adalah salah satu metode dalam data mining yang bertujuan untuk memprediksi kategori atau label dari data berdasarkan atribut yang tersedia. Teknik ini banyak digunakan dalam berbagai aplikasi seperti deteksi spam, prediksi penyakit, atau pengelompokan pelanggan.

- Pengertian Ensemble Method

Ensemble method adalah teknik yang menggabungkan prediksi dari beberapa model untuk meningkatkan akurasi dan kestabilan hasil prediksi. Dengan menggabungkan beberapa model, kelemahan dari model individu dapat diminimalkan.

Jenis-Jenis Ensemble Method

1. Bagging (Bootstrap Aggregating):

- Membuat beberapa model dari subset data yang berbeda (hasil sampling dengan

pengembalian).

- Model individu (seperti decision tree) dilatih secara terpisah, dan prediksi akhir diperoleh dengan rata-rata (untuk regresi) atau voting (untuk klasifikasi).

- Contoh: Random Forest.

2. Boosting:

- Melatih model secara berurutan, di mana setiap model baru difokuskan untuk memperbaiki kesalahan dari model sebelumnya.

- Contoh: AdaBoost, Gradient Boosting, XGBoost.

3. Stacking:

- Menggabungkan beberapa model dengan melatih model meta (model tingkat kedua) untuk membuat prediksi akhir berdasarkan prediksi dari model awal.

- Keuntungan Ensemble Method

- Mengurangi risiko overfitting.

- Meningkatkan akurasi dibandingkan model individu.

- Dapat menangani data yang kompleks dan tidak seimbang.

- Pengertian Class Imbalance

Class imbalance terjadi ketika distribusi kelas dalam dataset tidak seimbang, yaitu salah satu kelas memiliki jumlah data yang jauh lebih banyak dibandingkan kelas lainnya. Contohnya, dalam dataset pendeteksian penipuan, jumlah transaksi normal jauh lebih banyak daripada transaksi penipuan.

- Masalah Akibat Class Imbalance

1. Model cenderung bias terhadap kelas mayoritas.

2. Penurunan akurasi dalam mendeteksi kelas minoritas.

3. Metrik seperti akurasi menjadi kurang relevan karena tidak mencerminkan performa model pada kelas minoritas.

- Pendekatan untuk Mengatasi Class Imbalance

1. Teknik Sampling:

- Oversampling: Menambahkan data kelas minoritas dengan teknik seperti SMOTE (Synthetic Minority Oversampling Technique).

- Undersampling: Mengurangi jumlah data dari kelas mayoritas untuk menyeimbangkan distribusi.

2. Penyesuaian Algoritma:

- Menggunakan algoritma yang dirancang untuk menangani data tidak seimbang, seperti cost-sensitive learning.

3. Penggunaan Metrik yang Sesuai:

- Gunakan metrik seperti F1-score, precision, recall, atau AUC-ROC yang lebih mencerminkan performa pada kelas minoritas.

4. Ensemble Method:

- Menggabungkan beberapa model dengan teknik seperti boosting yang dirancang untuk menangani kesalahan pada kelas minoritas.

- Contoh Aplikasi

1. Deteksi Penipuan (Fraud Detection):

- Dataset transaksi keuangan sering mengalami class imbalance, dengan jumlah transaksi normal yang jauh lebih banyak dibandingkan transaksi penipuan.

- Teknik yang digunakan: SMOTE untuk oversampling, Random Forest, atau Gradient Boosting untuk model klasifikasi.

2. Diagnosis Penyakit Langka:

- Jumlah pasien dengan penyakit langka sering kali jauh lebih sedikit dibandingkan pasien normal.

- Teknik yang digunakan: Cost-sensitive learning, SMOTE, dan ensemble methods.

Panduan Praktikum

Download lalu pelajari dan running code pada link berikut:
https://github.com/aldinata/Modul-Praktikum-Data-Mining/blob/main/Materi/3%20-%20Classification%20(Ensemble%20Method%20%26%20Class%20Imbalance).ipynb