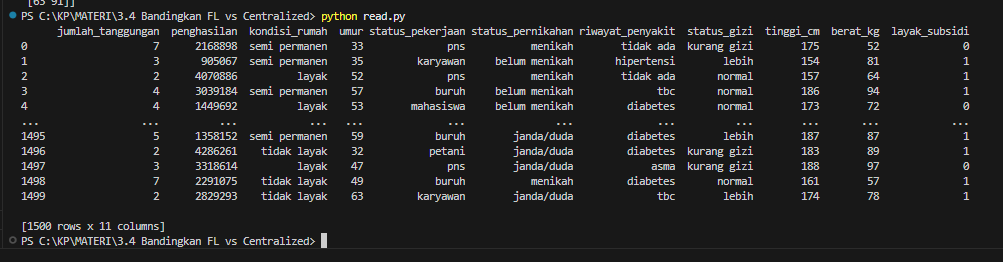
Topik : 3.4. Bandingkan FL vs Centralized

Objective : Latih model centralized untuk dataset gabungan → bandingkan

Task : Buat tabel perbandingan akurasi & risiko privasi

Model centralized dengan Scikit-learn berarti kita melatih model machine learning secara terpusat (centralized learning), di mana semua data dari berbagai sumber dikumpulkan dulu dalam satu tempat/server, baru kemudian dipakai untuk training.

Gabungan CSV :



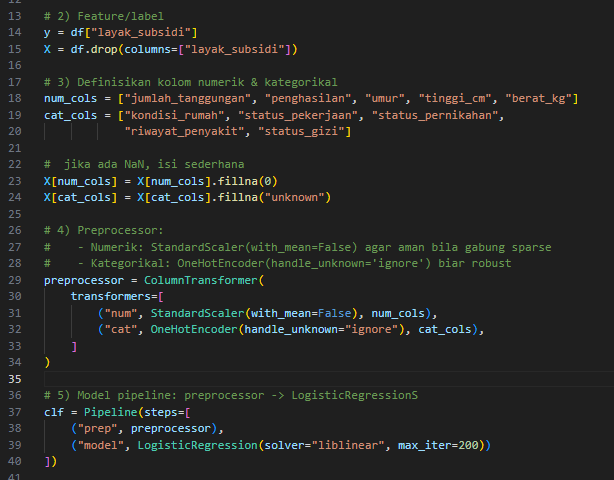
Centralized :

1. Load Dataset



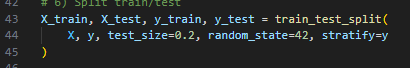
Semua data dari berbagai sumber (Dinsos, Dukcapil, Kemenkes) dikumpulkan jadi satu file CSV gabungan.

1. Preprocessing



Handle missing value, scaling fitur numerik agar model lebih stabil.

1. Split Data



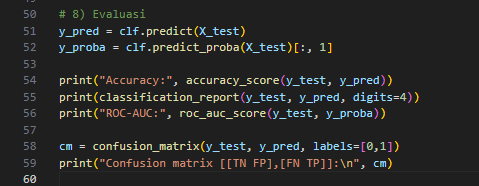
Data dipisah jadi training (80%) dan testing (20%).

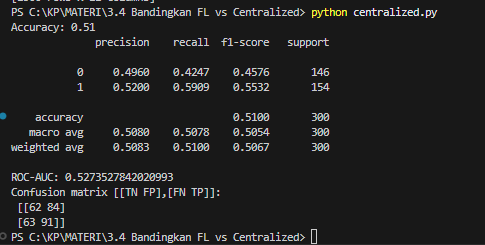
1. Train centralized model



Model dilatih dengan seluruh data

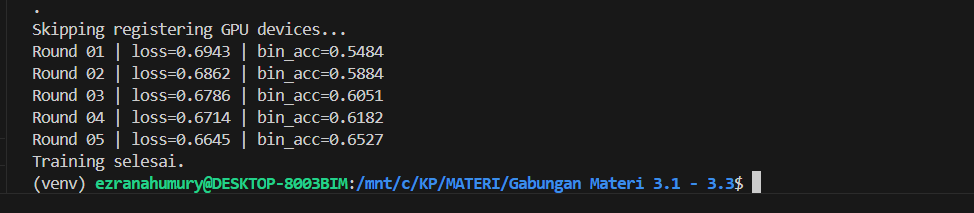
1. Evaluasi



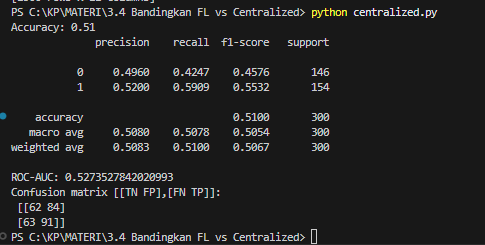


Ukur performa model dengan accuracy, precision, recall, dan f1-score.

Hasil FL :



Hasil Centralized :



**Tabel Perbandingan FL vs Centralized**

|  |  |  |
| --- | --- | --- |
| Aspek | Federated Learning (FL) | Centralized Learning |
| Accuracy | |  | | --- | |  |  |  | | --- | | 0.54 → 0.65 setelah 5 round (bin\_acc) | | ±0.51 (ROC-AUC 0.52, hampir random) |
| Data | Data tetap di client, hanya model update yang dibagi | Semua data harus digabung di server |
| Privasi | Lebih aman: data mentah tidak keluar dari perangkat | Risiko tinggi: semua data dikirim & disimpan |
| Risiko Privasi | Rendah → karena raw data tetap lokal | Tinggi → data sensitif rentan bocor |
| Kelebihan | Aman untuk data sensitif (kesehatan, finansial, dsb.) | Sederhana, cepat, library banyak tersedia |

1. **Performa Model**

Dari hasil percobaan, Federated Learning menunjukkan akurasi yang cenderung naik dari 0.54 → 0.65 selama 5 ronde training, sementara model Centralized (Scikit-learn Logistic Regression) hanya mencapai akurasi sekitar 0.51 dengan ROC-AUC mendekati 0.52 (hampir setara random guess).

* Rendahnya akurasi centralized lebih disebabkan karena dataset dummy kita diberi label acak, sehingga sulit bagi model menemukan pola yang bermakna.
* FL terlihat sedikit lebih stabil karena sifat agregasi model antar-klien membuat hasil tidak terlalu bergantung pada distribusi data tunggal.

1. **Privasi Data**

* Centralized learning mengharuskan semua data dikumpulkan ke satu server. Hal ini berisiko tinggi untuk privasi, terutama jika data berasal dari instansi berbeda (contoh: Dinsos, Dukcapil, Kemenkes). Data mentah bisa terekspos jika server diretas atau ada kebocoran.
* Federated Learning menjaga data tetap berada di perangkat atau institusi masing-masing. Hanya parameter model (gradien atau bobot terupdate) yang dikirim ke server pusat. Risiko kebocoran data mentah jauh lebih kecil, menjadikannya lebih cocok untuk kasus sensitif seperti kesehatan, keuangan, atau data kependudukan.