Opis Matematyczny Podziału Danych

Federated Learning z Stratyfikowaną 10-Fold Cross-Validation

1. Symbole matematyczne

Symbol	Opis
D	Całkowity dataset (zbiór wszystkich danych)
|D|	Liczba wierszy w datasecie D
D_tr	Zbiór danych treningowych (80%)
D_val	Zbiór danych walidacyjnych (10%)
D_ts	Zbiór danych testowych (10%)
F_i	i-ty fold treningowy (i = 1, 2, ..., 10)
C_k	k-ty klient (k = 1, 2, ..., K)
K	Liczba klientów (2, 3, 4 lub 5)
y	Zmienna docelowa (klasy)
c	Klasa (c należy do {c1, c2, ..., cm})
p(c)	Proporcja klasy c w zbiorze

2. Główny podział danych: Stratyfikowany 80/10/10

Formuła podziału:

$$D = D_{tr} \cup D_{val} \cup D_{ts}$$

gdzie:

- $|D_{tr}| = 0.80 \times |D|$ - dane treningowe
- $|D_{val}| = 0.10 \times |D|$ - dane walidacyjne
- $|D_{ts}| = 0.10 \times |D|$ - dane testowe

Warunek stratyfikacji:

Dla każdej klasy c proporcje są zachowane we wszystkich zbiorach:

$$p(c \mid D_{tr}) = p(c \mid D_{val}) = p(c \mid D_{ts}) = p(c \mid D)$$

Właściwości:

- D_tr, D_val, D_ts są rozłączne (żadne dane się nie powtarzają)
- Stratyfikacja gwarantuje identyczne proporcje klas we wszystkich zbiorach

3. Stratyfikowana 10-Fold Cross-Validation

Podział danych treningowych na 10 foldów:

$$D\_tr = F\_1 \cup F\_2 \cup F\_3 \cup ... \cup F\_10$$

Rozmiary foldów:

$$|F\_i| = |D\_tr| / 10 \quad \text{dla } i = 1, 2, ..., 10$$

Warunek stratyfikacji foldów:

Dla każdej klasy c i każdego foldu F_i:

$$p(c \mid F\_i) = p(c \mid D\_tr) = p(c \mid D)$$

Każdy fold ma identyczny rozkład klas w przybliżeniu

4. Mechanizm Cross-Validation

Dla każdej iteracji CV (i = 1, 2, ..., 10):

- **Dane treningowe w iteracji i:** D_train(i) = wszystkie foldy OPRÓCZ F_i (9 foldów)
- **Dane walidacyjne lokalne w iteracji i:** D_val_local(i) = F_i (1 fold)

Iteracja	Trening (9 foldów)	Walidacja lokalna
i = 1	F_2 ∪ F_3 ∪ ... ∪ F_10	F_1
i = 2	F_1 ∪ F_3 ∪ ... ∪ F_10	F_2
i = 3	F_1 ∪ F_2 ∪ F_4 ∪ ... ∪ F_10	F_3
...	...	...
i = 10	F_1 ∪ F_2 ∪ ... ∪ F_9	F_10

5. Podział na klientów (Federated Learning)

Rotacja foldów:

Każdy klient otrzymuje te same 10 foldów, ale w różnej kolejności:

$$\text{start\_fold}(k) = ((k - 1) \times \text{floor}(10 / K)) \text{ MOD } 10$$

Przykład dla K=2 klientów:

Klient	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Client 1	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Client 2	F6	F7	F8	F9	F10	F1	F2	F3	F4	F5

Współdzielone zbiory:

- D_val - identyczny dla wszystkich klientów (SHARED)
- D_ts - identyczny dla wszystkich klientów (SHARED)

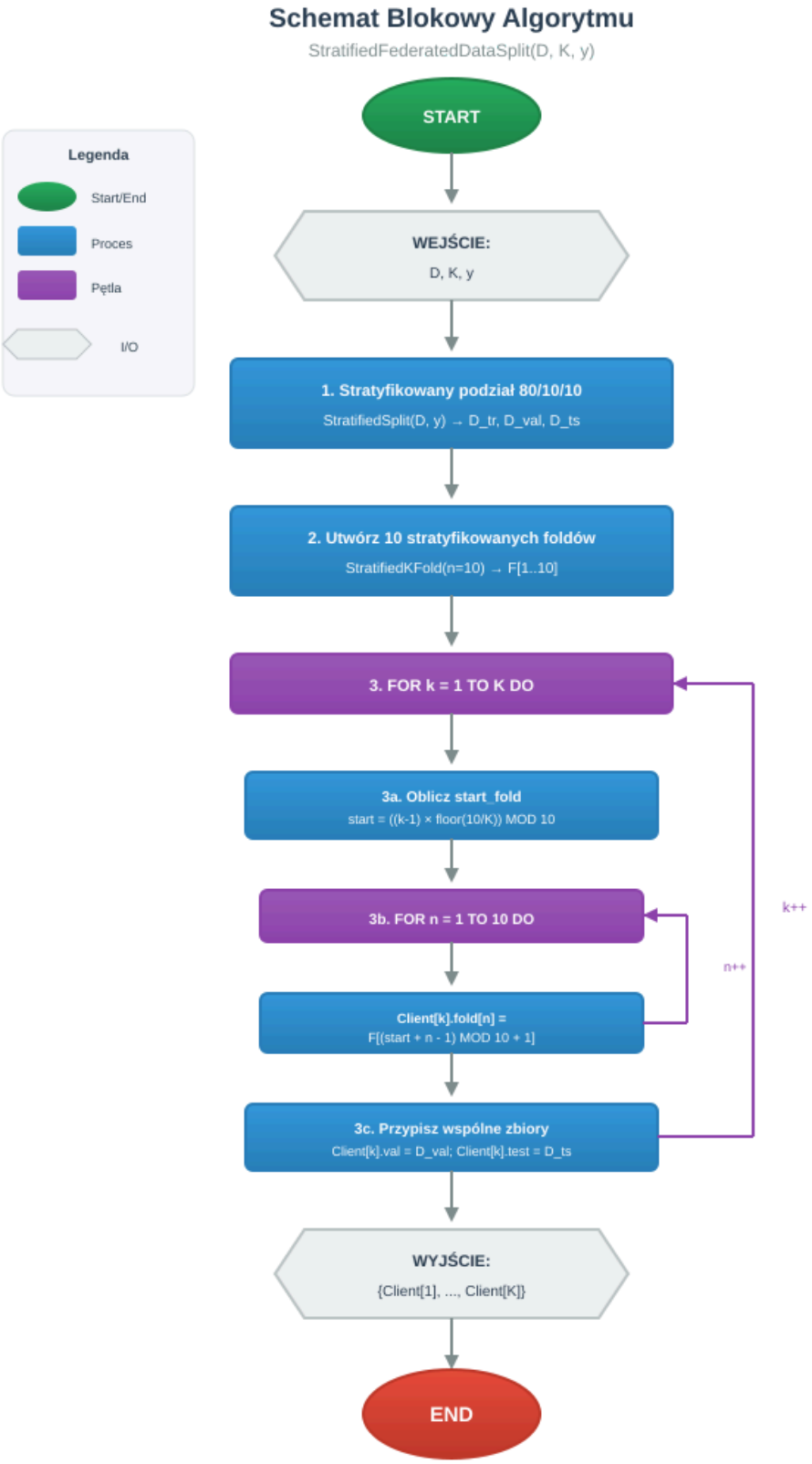
6. Algorytm podziału danych ze stratyfikacją

ALGORYTM: StratifiedFederatedDataSplit(D, K, y)

WEJŚCIE: D (dataset), K (liczba klientów), y (zmienna docelowa)
WYJŚCIE: K klientów z foldami, val, test

```
BEGIN
  1. // STRATYFIKOWANY PODZIAŁ 80/10/10
     [D_train, D_temp] = StratifiedSplit(D, y, test_size=0.20)
     [D_val, D_test] = StratifiedSplit(D_temp, y_temp, test_size=0.50)
  2. // STRATYFIKOWANE FOLDY (StratifiedKFold)
     skf = StratifiedKFold(n_splits=10, shuffle=True)
     FOR i = 1 TO 10 DO Fold[i] = D_train[skf.indices[i]]
  3. // PRZYPISANIE DO KLIENTÓW
     FOR k = 1 TO K DO
       start_fold = ((k-1) × floor(10/K)) MOD 10
       Client[k].fold[n] = Fold[(start_fold + n - 1) MOD 10 + 1]
       Client[k].val = D_val; Client[k].test = D_test
  4. RETURN {Client[1], ..., Client[K]}
END
```

7. Schemat blokowy algorytmu



8. Przykład: RT-IoT2022

Dane: |D| = 123,117 wierszy, K = 2 klientów

Obliczenia:

- |D_tr| = 0.80 × 123,117 = 98,493 wierszy
- |D_val| = 0.10 × 123,117 = 12,312 wierszy
- |D_ts| = 123,117 - 98,493 - 12,312 = 12,312 wierszy

Weryfikacja: 98,493 + 12,312 + 12,312 = 123,117

Rozmiar folda: |F_i| = 98,493 / 10 ≈ 9,849 wierszy

9. Podsumowanie wszystkich datasetów

Dataset	Total	Train (80%)	Val (10%)	Test (10%)	Folds	Stratyfikacja
RT-IoT2022	123,117	98,493	12,312	12,312	10	TAK
Letter Recognition	20,000	16,000	2,000	2,000	10	TAK
Electric Power	2,049,279	1,639,423	204,928	204,928	10	NIE*
MIMIC-IV-ED	196	156	20	20	10	TAK

**Electric Power to dane czasowe (regresja) - stratyfikacja nie ma zastosowania.*

10. Kluczowe koncepcje

Koncepcja	Opis
Stratyfikacja	Zachowanie identycznych proporcji klas w każdym zbiorze i foldzie
Disjointness	Zbiory D_tr, D_val, D_ts są rozłączne - żaden wiersz się nie powtarza
Completeness	D_tr ∪ D_val ∪ D_ts = D - każdy wiersz użyty dokładnie raz
Fold Rotation	Klienci mają te same dane, foldy w różnej kolejności
Shared Val/Test	Wszystkie klienty używają identycznych D_val i D_ts
Cross-Validation	10 iteracji: fold F_i jako walidacja lokalna, pozostałe 9 jako trening

11. Wzory matematyczne

Całkowity rozmiar danych dla K klientów:

$$\text{Total} = K \times |D_{tr}| + |D_{val}| + |D_{ts}| = (0.80K + 0.20) \times |D|$$

Warunek stratyfikacji (dla każdej klasy c):

$$\text{Dla każdego } c\text{: } |p(c|F_i) - p(c|D)| < \varepsilon, \text{ gdzie } \varepsilon \to 0$$

Rozmiar danych w jednej iteracji CV:

$$|D_{train(i)}| = |D_{tr}| - |F_i| = 0.9 \times |D_{tr}|$$

$$|D_{val\_local(i)}| = |F_i| = 0.1 \times |D_{tr}|$$