프로젝트 계획: 친족 관계 분류

목표: IBD 핵심 지표와 `merged_info.out`에서 파싱한 분포 통계를 결합해 다중 클래스 친족 관계를 예측하고, 데이터셋별(cM_1, cM_3, cM_6) 통합 리포트를 생성합니다.

1단계: 데이터 준비 (각 cM 데이터셋)

1. `merged_info.out` 처리

- 압축 해제 후 각 라인을 파싱해 분포 통계(평균, 표준편차, 분위수 등)를 컬럼으로 구성
- `pair` 단위 한 줄(DataFrame 한 행)로 정리

2. `model_input_with_kinship_filtered_<dataset>.csv` 처리

- 필수 컬럼만 사용: `pair`, `IBD1_len`, `IBD2_len`, `R1`, `R2`, `Num_Segs`, `Total_len`, 타깃 `kinship`

3. 병합 및 시나리오 분기

- `pair` 기준 병합 후 두 시나리오 구성:
  - included: UN 포함
  - noUN: `kinship=='UN'` 제거
- 산출물(재생성 가능, git 무시):
  - `data/processed/merged_<dataset>.csv`
  - `data/processed/merged_<dataset>_noUN.csv`

2단계: EDA

- 타깃 분포(kinship) 막대 그래프 생성 → 클래스 불균형 점검

3단계: 특성 선택 & 전처리

- X = IBD 6개 + 분포 통계 전체, y = kinship
- `StandardScaler` 적용 (선택된 특성에 대해 저장)
- `RandomForestClassifier` 중요도로 상위 특성(예: 50개) 선택 및 시각화

4단계: 모델 학습 (CUDA 전용)

- 불균형 전략:
  - zero: 재균형 없음
  - weighted: 클래스 가중 손실
  - smote: 학습 세트 과샘플링
  - overunder: SMOTE 후 ENN/Tomek으로 경계 노이즈 제거
- 모델:
  - 고급 MLP (BatchNorm/Dropout 포함 깊은 구조)
  - 고급 1D-CNN (3개 Conv 블록 + 2개 FC)
- 특수 학습 스케줄: UN 포함 + 과샘플링(smote/overunder)일 때만 `--special-epochs` 적용

5단계: 평가 & 리포팅

- 지표: Accuracy, F1(가중/매크로), AUC(가중/매크로/마이크로; OvR, N/A 없음), 혼동 행렬
- 클래스 불균형 참고:
  - zero는 기준선(다수 클래스 편향 가능)
  - weighted는 클래스 가중, smote는 학습만 과샘플링(검증 분포는 원본)
  - overunder는 과샘플 후 경계 정리로 경계 선명화 기대
- 리포트: 시나리오 플롯(분포/특성중요도)은 SVG/PNG로 저장, 본문은 2열 배치; 혼동 행렬 포함

6단계: 반복 실행

- 각 데이터셋(cM_1, cM_3, cM_6)에 대해 두 시나리오와 모든 전략(zero/weighted/smote/overunder) 실행

- 산출물: `reports/<dataset>/` 에 데이터셋별 단일 통합 리포트(영문/국문 Markdown, 선택적 PDF)

연구/후속 과제

1. 과샘플링 유무 비교
2. UN 라벨 영향(시나리오 분리 + 샘플링 전략) 분석
3. 비과샘플링 러닝에서 에포크 증가 효과 검토
4. 상위 중요도 특성의 추가 통계 파생 검토
5. overunder(ENN vs Tomek) 변형별 성능/강건성 비교