OUTLINE [B팀]

팀원: 김민섭, 노연경, 송린, 안도원

1. 주제 선정 배경 및 목적

- 의료 분야에서 AI 기반 조기 진단 보조 모델의 중요성
- 갑상선암 조기 판별을 통한 환자 관리 효율화 및 진단 정확도 제고

2. 데이터 출처 및 간단한 설명

- 데이콘 갑상선암 진단 데이터
- train.csv, test.csv 두가지 파일로 이루어져있음

컬럼명	데이터 타입	설명	값(범주 / 범위)
ID	string	샘플별 고유 ID	`TRAIN_00000` ~ `TRAIN_87158`
Age	integer	환자의 나이	14 ~ 88
Gender	string	성별	`M`, `F`
Country	string	국적 (ISO 3자리 코드)	예: `CHN`, `NGA`, `IND`, `USA`, `GBR` 등
Race	string	인종 코드	`ASN`, `MDE`, `HSP`, `CAU`, `AFR` 등
Family_Background	string	가족력 여부	`Positive`, `Negative`
Radiation_History	string	방사선 노출 이력	`Exposed`, `Unexposed`
Iodine_Deficiency	string	요오드 결핍 여부	`Sufficient`, `Deficient`
Smoke	string	흡연 여부	`Non-Smoker`, `Smoker`
Weight_Risk	string	체중 관련 위험도	`Not Obese`, `Obese`
Diabetes	string	당뇨병 여부	`No`, `Yes`
Nodule_Size	float	갑상선 결절 크기 (cm)	0.0 ~ 5.0
TSH_Result	float	TSH 호르몬 검사 결과 (μIU/mL)	0.1 ~ 10.0
T4_Result	float	T4 호르몬 검사 결과 (μg/dL)	4.5 ~ 12.0
T3_Result	float	T3 호르몬 검사 결과 (ng/mL)	0.5 ~ 3.5
Cancer	integer	갑상선암 여부 (타깃)	0 (양성), 1 (악성)

sample

ID	Age	Gender	Country	Race	Family_Background	Radiation_History	Iodi
TRAIN_000000	80	M	CHN	ASN	Positive	Exposed	Suf
TRAIN_000001	37	M	NGA	ASN	Positive	Unexposed	Suf
TRAIN_000002	71	M	CHN	MDE	Positive	Unexposed	Suf

3. 해결하고자 하는 문제 정의

- **문제 유형**: 이진 분류(Binary Classification)
- **목표**:
  1. 환자의 정형 건강·생활 데이터를 바탕으로 '양성(0)'인지 '악성(1)'인지를 정확히 예측
  2. 특히 악성 종양을 놓치는 **False Negative**를 최소화하면서도, F1 score를 최적화하는 모델 개발
- **평가 지표**:
  - **Primary**: F1 Score
  - **Secondary**: Precision, Recall, ROC-AUC 등
- **프로젝트 계획**:

```
flowchart TD
    A["데이터 전처리<br/>EX) 결측치 처리, 범주형<br/>인코딩, 스케일링"] --> B["탐색적 데이터 분석 (EDA)"]
    B --> C["모델 학습 및 하이퍼파라미터 튜닝"]
    C --> D["모델 선택<br/>테스트 세트 예측"]
    D --> E["평가 결과 분석<br/>모델 해석 및 보고서 작성"]
```