# 1. 예측 유지보수: GPU 가속 XGBoost를 활용한 시계열 분류

## 프로젝트 개요

Backblaze의 하드 드라이브 SMART 데이터를 사용하여 고장/정상 상태를 분류하는 기계 학습(ML) 기반의 예측 유지보수(Predictive Maintenance) 모델을 구축했습니다. 특히, **NVIDIA RAPIDS** 라이브러리를 활용하여 **XGBoost 모델의 학습 및 테스트 과정**을 GPU로 가속하여 성능을 최적화했습니다.

## ⚙️ 주요 기술 및 역량

* **기술:** NVIDIA RAPIDS(cuDF, cuML), XGBoost Classifier, 시계열 데이터 전처리.
* **핵심 역할:**
  1. **데이터 준비 및 정제:** 방대한 시계열 데이터셋에서 고장 예측에 불필요한 컬럼을 제거하고, 누락된 값(NaN/Null)을 처리하는 데이터 큐레이션 과정을 수행했습니다.
  2. **클래스 불균형 해소:** 하드 드라이브 고장 데이터(Faulty)가 전체 데이터의 극히 일부만을 차지하는 **클래스 불균형(Imbalanced Classes)** 문제를 해결하기 위해 데이터 오버샘플링(Oversampling) 또는 가중치 부여 기법을 적용했습니다.
  3. **GPU 가속:** cuDF 데이터 구조를 사용하여 XGBoost 모델을 훈련함으로써, 대규모 데이터셋에서도 **CPU 대비 월등히 빠른 속도**로 모델을 훈련하고 평가했습니다.

## 🛠️ 주요 구현 내용

1. **cuDF 로드:** Pandas 대신 cuDF를 사용하여 데이터를 GPU 메모리에 직접 로드하고, 모든 전처리 작업을 GPU 상에서 처리했습니다.
2. **특징 공학(Feature Engineering):** 시간 흐름에 따른 고장 예측을 위해 필요한 핵심 SMART 지표만을 선별하고 정규화(Normalization)했습니다.
3. **모델 훈련:** cuML 라이브러리의 GPU 기반 XGBoost 모델을 사용하여 분류 모델을 훈련하고, 테스트 데이터셋에서 모델의 정확도와 성능 지표(AUC-ROC)를 검증했습니다.

## 📌 주요 코드 로직 (GPU 기반 데이터 처리 및 모델 훈련 추상화)

import cudf  
import cuml.ensemble as cuml\_xgb  
  
# 1. 데이터 로드 및 전처리 (GPU 가속)  
df\_cudf = cudf.read\_csv('...')  
  
# 2. 클래스 불균형 처리 (예: 가중치)  
  
# 3. cuML XGBoost 모델 훈련  
model = cuml\_xgb.XGBoostClassifier(tree\_method='gpu\_hist')  
model.fit(X\_train\_cudf, y\_train\_cudf)  
  
# 4. 예측  
y\_pred = model.predict(X\_test\_cudf)