실습: 편향 테스트 코드 (Hands-on: Bias Testing)

Python 편향 테스트 예제

```
from sklearn.metrics import confusion_matrix import pandas as pd # 그룹별 성능 평가 def
evaluate_fairness(y_true, y_pred, sensitive_attr): results = {} for group in
sensitive_attr.unique(): mask = sensitive_attr == group tn, fp, fn, tp = confusion_matrix(
y_true[mask], y_pred[mask] ).ravel() results[group] = { 'TPR': tp / (tp + fn), # Recall 'FPR': fp
/ (fp + tn), 'PPV': tp / (tp + fp) # Precision } return pd.DataFrame(results).T
```

테스트 도구

- Fairlearn (Microsoft)
- AI Fairness 360 (IBM)
- What-If Tool (Google)
- Aequitas

시각화

- 그룹별 성능 비교 차트
- Confusion matrix 히트맵
- Fairness metric 대시보드
- ROC curve 비교

실습 포인트

다양한 인구 집단에 대해 TPR, FPR, PPV를 계산하고 격차가 허용 범위(예: 10%) 내인지 확인