우선 문제는 사전공개했듯이
관세청에서 수입품의 정보를 보고 우범(불법)인지 아닌지를 예측하는 문제입니다.
1-12월 데이터가 있는데
1-9월을 학습데이터로 사용하고, 10-12월을 평가데이터로 사용합니다.

1. train.xlsx: 학습 데이터(y값 존재)
1-9월의 데이터로 가장 오른 쪽의 검사결과코드, 우범여부, 핵심적발이 종속변수인데
그 중에서 우범여부가 우선적으로 예측해야할 target입니다.

2. test.xlsx: 평가 데이터(y값 없음)
10-12월의 데이터로 3개의 종속변수가 삭제되었습니다.

3. submission.xlsx: 제출용 평가 sheet
여기 신고번호에 맞게 예측값을 붙여서 제출하시면 됩니다.
단 제출하실 때 파일명을 팀명으로 해주세요.

- 우범여부 예측이 1차적인 본 과제의 과업이지만
좀 더 advanced로 가실 분들은 우범여부와 더불어 핵심적발여부도 해보시면 좋을 것 같습니다.

- 정상:비정상 비율에서 정상이 훨씬 많은 범주 불균형 문제입니다.
일반적인 2-class 분류기로 접근하면 성능이 낮아질 수 있습니다.
범주 불균형 문제에 대해서는 이번 주 이론시간에 커버하도록 하겠습니다.

- 평가지표는 범주 불균형 문제에 맞게
precision, recall, f1 score 이 세가지를 기준으로 보시면 됩니다.
하나의 지표를 기준으로 한다면 f1을 우선적으로 봐주세요.

- 학습 데이터 내에서 validation을 충분히 하셔야
그 모델로 평가 데이터의 class label을 잘 달 수 있겠죠?
cross-validation, moving-window 등을 통해서 validation에 신경을 많이 써야할 것 같습니다.

- 나머지는 참고자료 및 참고코드입니다.
참고코드는 같은 문제이지만 다른 데이터셋을 기준으로 작성이 되었으므로
이번 데이터에 맞게 돌려보시면서 방향을 결정하시면 될 것 같습니다.

- 참고코드를 최대한 활용하시되 그대로 하시면 안됩니다.
아이디어만 참고하시고 새로운 방법을 해보세요.
이걸 그대로 쓰시면 아무래도 나중에 좋은 점수를 드리기 어려울 것 같습니다.