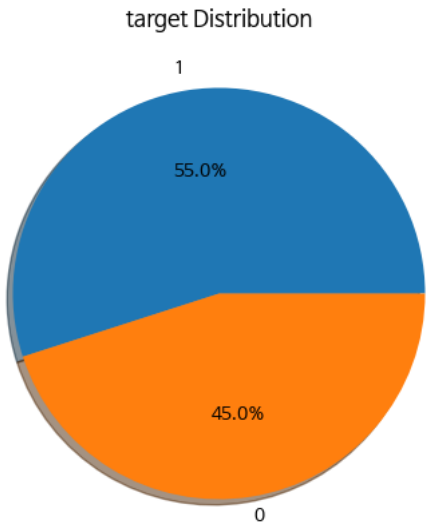
1. 데이터 불러오기
2. EDA
    A. Target의 분포 확인
    B. Features 분포 확인
3. 데이터 전처리
    A. 이상치를 평균으로 대체
    B. 연속형 변수를 표준화
4. 모델 학습
    A. 기본적인 RandomForest 모델 생성
    B. GridsearchCV를. 통한 최적의 파라미터 찾기
5. 제출용 파일 생성

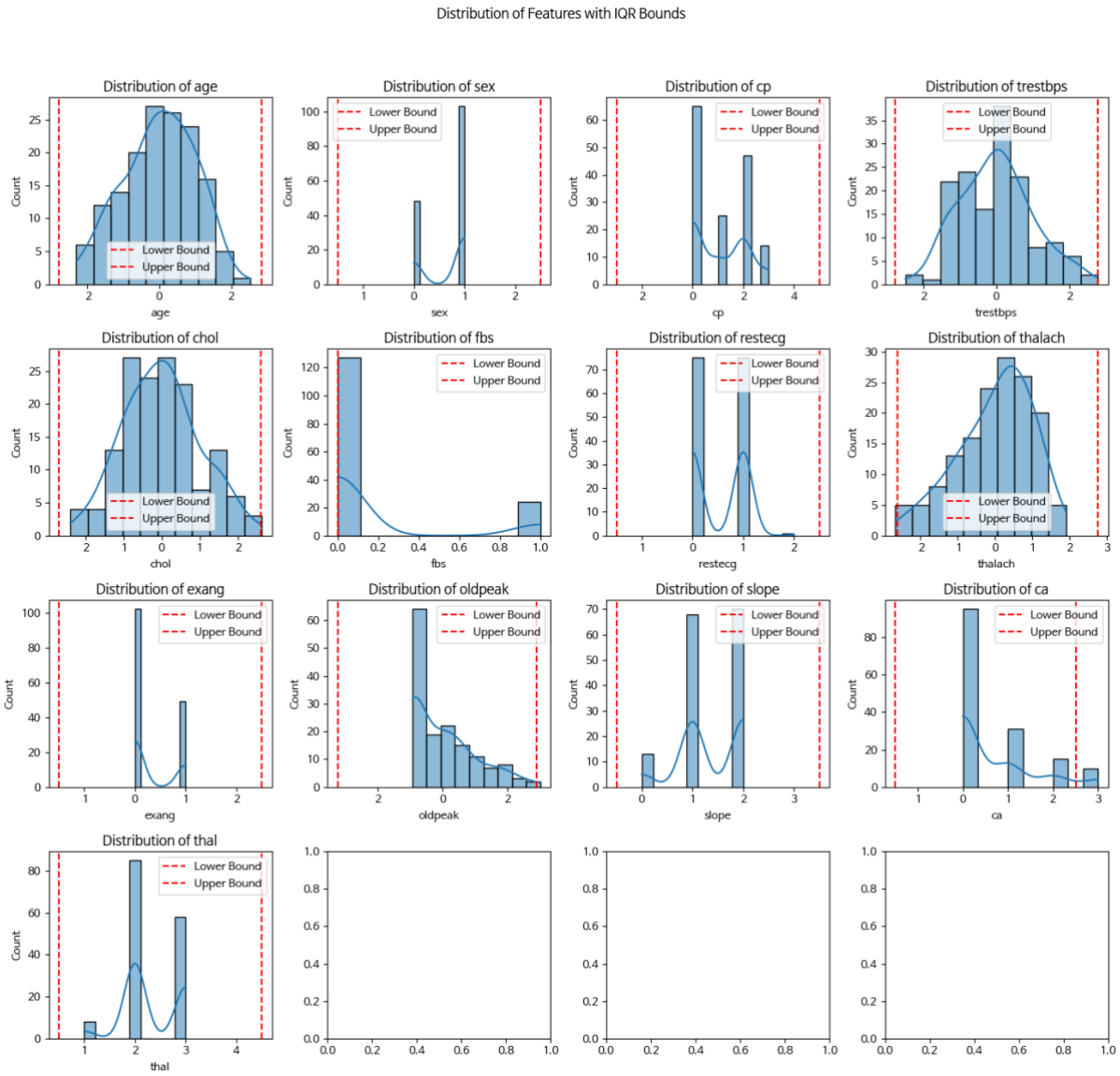
2-A.. Target의 분포 시각화



2-B. Features 분포 확인

분포를 확인하기 전 결측치에 대한 처리를 하였다.

ca=4, thal=0인 값은 결측치이므로, 중앙값으로 전부 대체 후 분포를 확인하였다.



3-A. 이상치 처리

trestbps, oldpeak, chol, ca에서 이상치가 발견되었으나, ca의 값이 3인 경우는 IQR 기준으로 이상치에 해당하지만, 이는 실제 혈관 수가 3개임을 의미하므로 이상치로 간주하기 어렵다. 따라서 ca를 제외한 컬럼

의 이상치를 평균으로 대체하였다.

3-B. 연속형 변수 표준화

성별과 같은 범주형 변수들은 이미 0과 1로 인코딩 되었기 때문에 연속형 변수에 대해서만 표준화 진행

4. 모델링

train set과 validation set에 대해 동일한 클래스 비율이 유지되도록 분할한 후, RandomForest 모델로 학습 진행을 하였으나 f1점수가 낮아서 GridSearchCV를 통해 RandomForest 모델의 최적의 하이퍼파라미터 탐색(3*3*3*2 -> 52가지 조합)

최적의 파라미터로 최종 학습 진행을 하였다.