[전처리 과정]

필요없는 컬럼 제거
의미가 없는 컬럼(이름, 날짜 등)을 제거한다. (line, name, mold_name, time, date, registration_time)
데이터 종류의 수가 1개인 컬럼을 제거한다. (emergency_stop - On만 존재, 의미없음)
molten_volume과 heating_furnace는 어느정도 연관이 있는 것으로 확인되기 때문에, molten_volume 컬럼을 제거한다. (해당 컬럼의 영향은 heating_furnace에서 확인한다.)
passorfail에서 1개의 행이 결측되어있어 해당 행을 제거한다.

결측치 채우기
tryshot_signal은 공정에 문제가 발생할 경우 발생하는 event로, 발생하지 않으면 따로 입력되지 않는 것으로 판단하여 결측치에 'No'를 채워주었다.
=> tryshot_signal 컬럼을 제외하고 학습을 진행한 결과 성능이 많이 떨어진다.

heating_furnace의 경우 50%의 결측치가 존재. 결측치의 대부분은 melten_volume과 일치하기 때문에 제가 3의 방식으로 보여진다. 즉, 결측치는 Other furance인 'C'로 채워서 분석을 진행하였다.
=> 아래의 상황에서 'A', 'B', 'C',' D'로 구분해서 모델 학습 시 약간 떨어지는 경향을 확인함.
=> melten_volume의 데이터가 있을 때 heating_furnace에 결측치가 있는 상황 ('C')
=> melten_volume의 데이터가 없을 때 heating_furnace도 결측된 상황 ('D')

upper_mold_temp3은 code별 최빈값으로 대체 (몰드코드 8412에서 결측 313행 존재)
=> 몰드코드 8573을 제외한 다른 몰드코드들은 데이터값이 1449만 존재,
=> 8573에서 1449값의 비율은 650/9500 수준으로 매우 낮아 최빈값을 사용해서 채워준다.

lower_mold_temp3은 code별 최빈값으로 대체 (몰드코드 8412에서 결측 313행 존재)
=> 몰드코드 8722을 제외한 다른 몰드코드들은 데이터값이 1449값만 존재,
=> 8722는 300~1449의 범위를 가지고 있으며, 1449의 비중은 18325/20004로 가장 높다.

molten_tem는 code별 최빈값으로 대체 (8412에서만 약 2천개의 결측치가 존재)
=> 대부분의 데이터값(약 19200개)이 700대의 값을 가지고 있으므로 최빈값으로 적용.

[데이터 인코딩]

'working' 컬럼 인코딩
```
df['working'] = df['working'].replace({'가동': 1, '정지': 0})
```

'heating_furnace' 컬럼 인코딩
```
df['heating_furnace'] = df['heating_furnace'].replace({'A': 0, 'B': 1, 'C': 2})
```

'mold_code' 컬럼 인코딩
```
df['mold_code'] = df['mold_code'].replace({8412: 0, 8413: 1, 8573: 2, 8576: 3,
                                           8600: 4, 8722: 5, 8917: 6})
```

[이상치 제거]

이상치를 제거하기 위해서 LocalOutlierFactor 기법을 적용하였다.
LOP를 적용한 이유 : 데이터의 수가 매우 많고(9만개 이상), 극단적인 이상치들이 존재하기 때문에 극단값들만 제거하기 위해서 LOF를 적용하였다.
극단적인 이상치 65503을(9개) 제거하기 위해서 기준은 n_neighbors=10으로 적용하였다.

[T-검정]

T-검정을 진행하여 pass(0)값과 fail값(1)에 대응하는 각 컬럼들 사이에서 t-검정을 진행하여 두 집단 사이에 '유의미한 차이'가 있는지 확인한다.
귀무 가설 (H0): 두 그룹(예: passorfail=1과 passorfail=0)의 평균이 동일하다.
대립 가설 (H1): 두 그룹의 평균이 다르다.
만약, t-검정 결과 p-값이 0.05보다 작다면 귀무 가설을 기각하고, "두 그룹의 평균에 유의미한 차이가 있다"라고 해석. 반대로 p-값이 0.05 이상이면 차이가 없다고 볼 수 있다.
즉, p-값이 0.05보다 작은 '유의미한 차이가 있는 컬럼'들을 사용하여 '0'과 '1'사이에 어떤 차이가 있는지 확인한다.
=> 실제 컬럼을 제거하고 진행한 것과 제거하지 않고 진행한 것은 차이가 거의 없었던 것으로 확인했음.