**㈜PCN**

**산업 데이터 증강**

**기술문서**

**Ver. 1.2**

**목 차**

1. **개발 환경**
   1. 개발언어
   2. IDE
   3. 라이브러리
   4. 프레임워크
2. **데이터 증강 절차**
   1. 설명
   2. 테스트 데이터
   3. 시스템 구성
3. **데이터 분포 확인** 
   1. Multi class 데이터 분포 확인
   2. Binary class 데이터 분포 확인
4. **데이터 전처리**
   1. 불필요한 Column 제거
   2. 증강할 수 없는 Target Data 제거
   3. Target Data 정수 인코딩
   4. 데이터 증강 비율 검증 및 설정
5. **데이터 증강 및 시각화**
   1. 데이터 증강
   2. 데이터 증강 결과 시각화
6. **개발환경**
   1. **개발 언어**
      1. Python 3.9
   2. **IDE**
      1. Vscode 1.83.1
   3. **라이브러리**
      1. json
      2. plotly
      3. Pandas
      4. Numpy
      5. Sklearn
      6. Imblearn
      7. requests
   4. **프레임워크**
      1. FastAPI
      2. streamlit
7. **데이터 증강 절차**
   1. **설명**

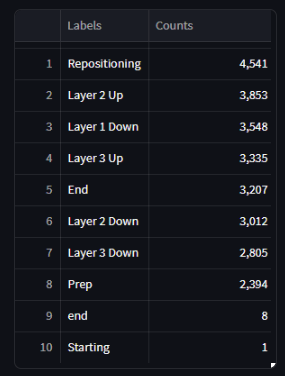
|  |  |
| --- | --- |
| **구분** | **설명** |
| 데이터 분포 확인 | * Target Data 확인 * 제거할 데이터 확인 * 정수 인코딩 필요한 데이터 확인 |
| 데이터 전처리 | * 불필요한 column, row 데이터 제거 * Target Data가 str 또는 object인 경우, 정수 인코딩 * Multi class, Binary class 옵션 선택 |
| 데이터 증강 및 시각화 | * SMOTE 방식으로 데이터 증강 * 증강 결과 Bar Chart, Pie Chart 시각화 * 증강 결과 테이블 확인 |

* 1. **데이터**
     1. Test\_01: CNC Miling machine
     2. Test\_02: Molding machine(사출성형기)
  2. **시스템 구성**

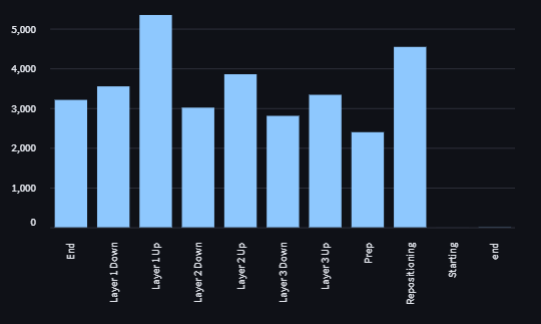


* + 1. 다중 분류인 경우, Streamlit으로 개발한 클라이언트에서 FastAPI로 개발한 서버에 Original data 전달 및 증강할 Target data 및 증강 비율 request.
    2. 이진 분류인 경우, 서버에 request하지 않고, Streamlit으로 개발한 클라이언트에서만 데이터 증강.

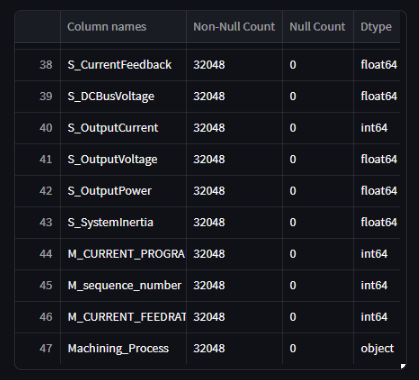
1. **데이터 분포 확인** 
   1. **Multi class 데이터 분포 확인**
      1. 데이터 증강할 Target Label 분포 테이블

.

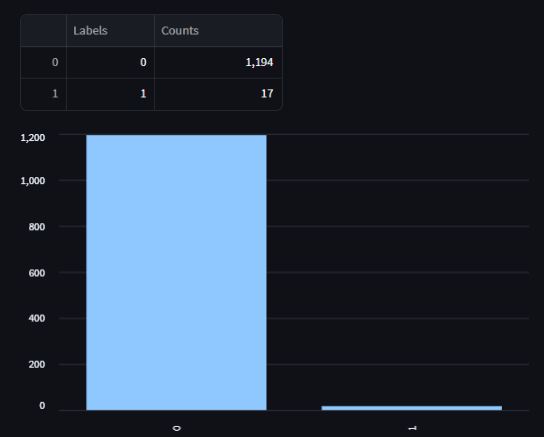
* + 1. 데이터 증강할 Target Label 분포 시각화.



* + 1. 데이터 증강해야 하는 Target Label의 테이블 및 시각화를 통해 제거할 데이터 확인.
    2. 데이터 증강 시 Target Label이 너무 적은 경우, 증강할 수 없기 때문에 데이터 전처리가 필요한지 확인. Ex) Target Label이 1개인 경우.

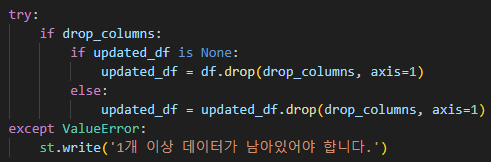


* + 1. 데이터 타입 확인
    2. Label Encoding이 필요한 object 타입의 Target Label 있는지 확인.
    3. object 또는 str 타입이 없다면, Label Encoding 하지 않음.
  1. **Binary class 데이터 분포 확인**
     1. 데이터 증강할 Target Label 분포 테이블 및 시각화

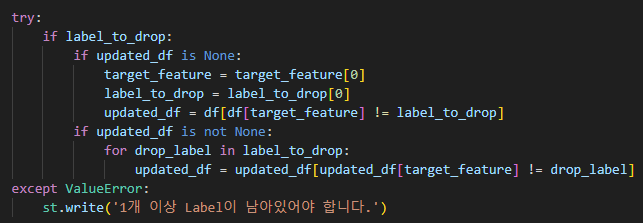


* + 1. Binary class의 경우, 기본 데이터 증강 비율은 2:1로 설정하지만, 지나치게 많은 차이가 있다면 데이터 증강 비율에 대한 threshold 조정이 필요한지 확인.

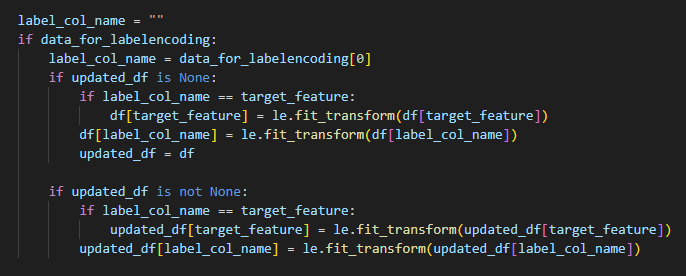
1. **데이터 전처리**
   1. **불필요한 Column 제거**



* + 1. 불필요한 인덱스 및 증강하지 않을 columns를 Pandas drop 함수를 사용하여 drop\_columns 변수에 선언되어 있는 column 제거
  1. **증강할 수 없는 Target Data 제거**

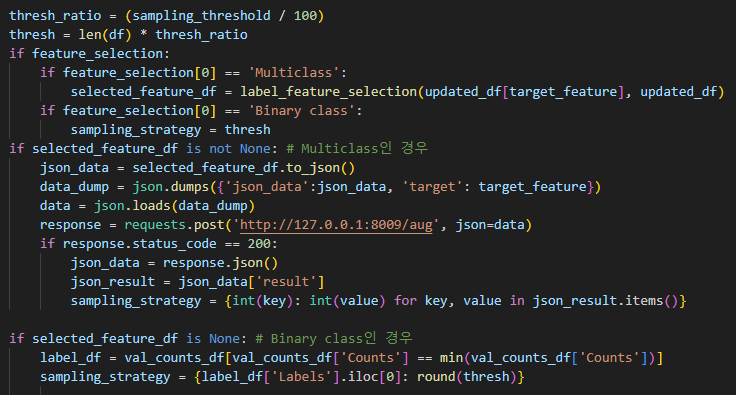


* + 1. drop\_label에 선언되어 있는 증강할 수 없는 Target Data를 제외하고 pandas DataFrame을 업데이트.
  1. **Target Data 정수 인코딩**

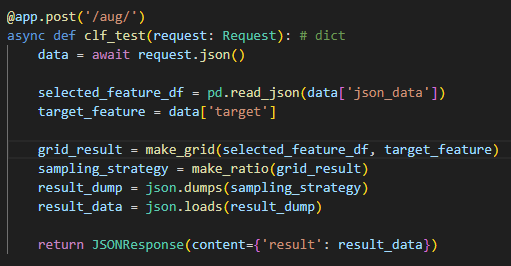


* + 1. Target Data의 타입이 str, object인 경우, 인식할 수 없기 때문에 int 타입으로 정수 인코딩.
    2. Target Data의 column과 Target Data가 아닌 column을 구분하고 scikit-learn의 LabelEncoder에 내장되어 있는 fit\_transform을 사용하여 Target Data에 대한 학습 및 인코딩을 동시에 진행.
  1. **데이터 증강 비율 검증 및 설정**

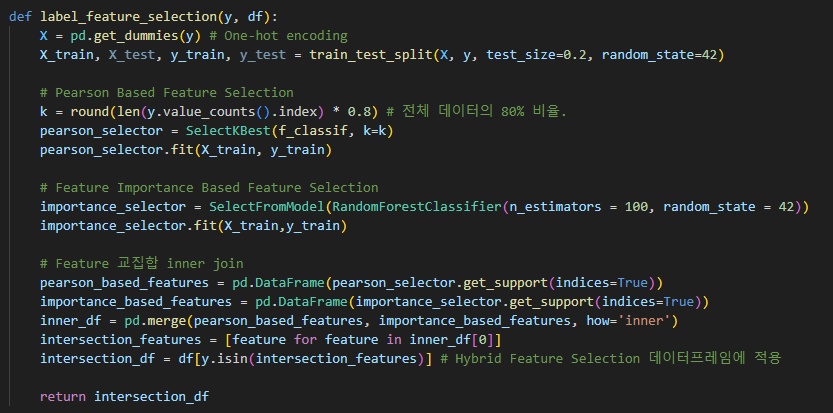




* + 1. Binary Class 데이터 증강 시 초기 threshold를 전체 데이터의 2:1 비율로 설정. Sampling\_threshold에 입력되는 threshold 값에 따라 데이터 증강 비율 변경.
    2. Multi class 데이터 증강 시, HFS(Hybrid Feature Selection)로 증강할 Target Data 선별 후, fastapi 서버에서 GridsearchCV를 사용하여 max, median, mean을 기준으로 데이터 증강 비율 검증 및 최적의 증강 비율 반환.
    3. Fastapi



* + - 1. 비동기 방식으로 클라이언트에서 전송되는 데이터를 받고 증강할 데이터 선별 및 증강 비율 측정.
      2. 최적의 증강 비율을 json으로 response.
    1. HFS(Hybrid Feature Selection)

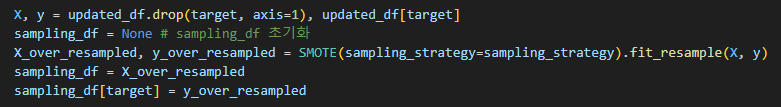


* + - 1. get\_dummies로 Target data를 One-hot encoding
      2. Pearson 방식과 Feature importance 방식의 교집합으로 Target data를 선별
    1. GridsearchCV

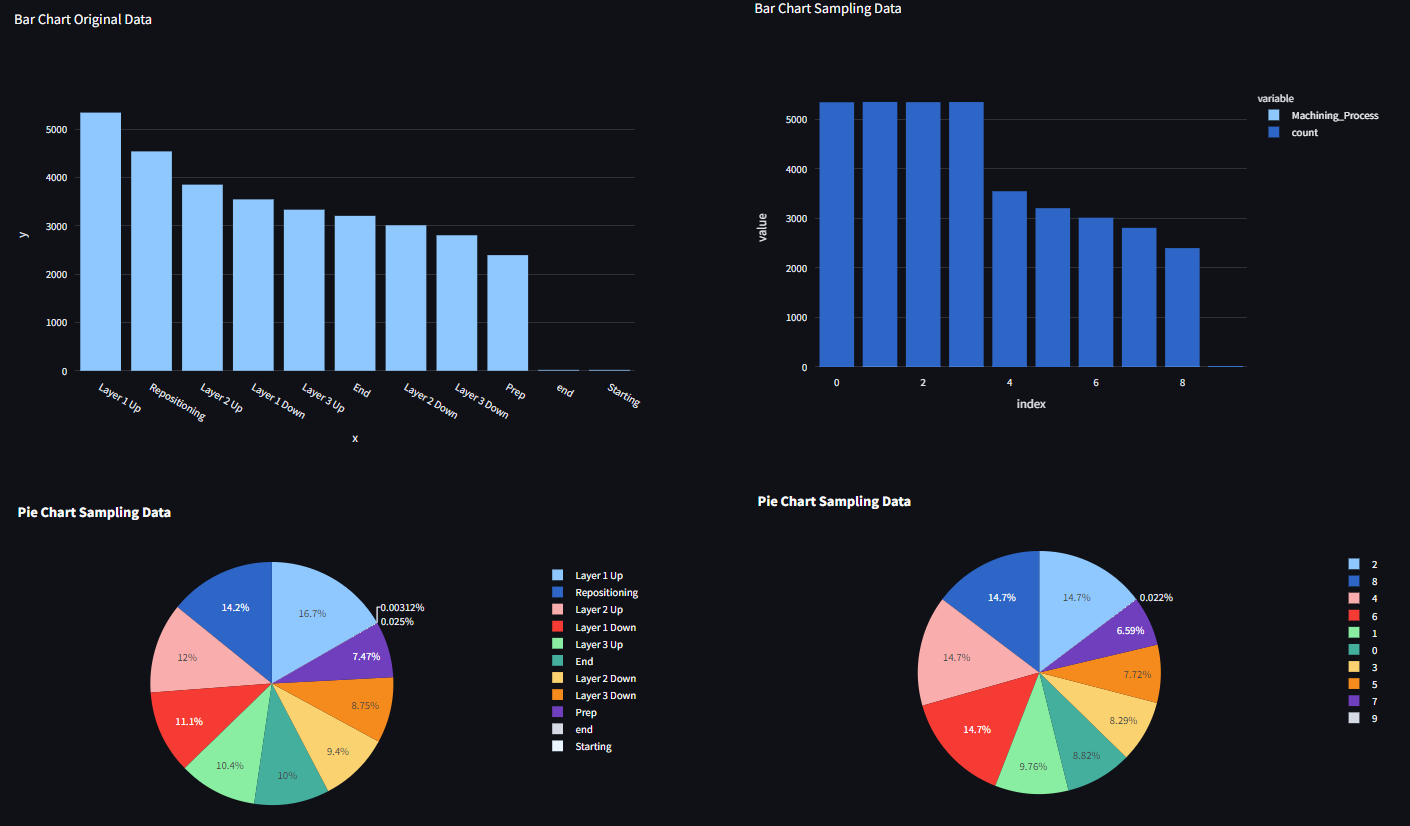


* + - 1. max, median, mean 값을 기준으로 순차적으로 증강 비율 검증

1. **데이터 증강 및 시각화**
   1. **데이터 증강**



* + 1. GridsearchCV로 설정한 데이터 증강 비율을 SMOTE 객체에 입력하여 데이터 증강.
    2. SMOTE 방식의 데이터 증강은 이웃한 데이터의 특성을 고려하여 데이터를 증강하는 기법.
  1. **데이터 증강 결과 시각화** 
     1. Streamlit을 사용하여 데이터 시각화



* + - 1. Bar chart에서 각 Target data의 수치 분포 확인
      2. Pie chart에서 각 Target data의 비율 분포 확인
    1. 데이터 증강 결과 출력



* + - 1. 테이블에서 증강한 데이터 확인