**㈜PCN**

**산업 데이터 모델 추천**

**기술문서**

**Ver. 1.2**

**목 차**

1. **개발환경**
   1. 개발언어
   2. IDE
   3. 라이브러리
   4. 프레임워크
2. **머신러닝 모델 추천 절차**
   1. 설명
   2. 테스트 데이터
   3. 시스템 구성
3. **데이터 분포 확인 및 시각화**
   1. Target Data 분포 확인
   2. 데이터 전처리가 필요한 데이터 Type 확인
4. **데이터 전처리**
   1. 불필요한 Columns 제거
   2. 불필요한 Target Data 제거
5. **AutoML을 활용한 머신러닝 모델**
   1. 지도학습 분류 모델
   2. 비지도학습 이상탐지 모델
6. **머신러닝 모델 추천** 
   1. 지도학습 기반 분류 모델 결과 확인
   2. 비지도학습 기반 이상탐지 모델 확인
7. **개발환경**
   1. **개발언어**
      1. python 3.9.13
   2. **IDE**
      1. Vscode
   3. **라이브러리**
      1. json
      2. ray
      3. pycaret
      4. pandas
      5. numpy
   4. **프레임워크**
      1. Streamlit
      2. FastAPI
8. **머신러닝 모델 추천 절차**
   1. **설명**

|  |  |
| --- | --- |
| **구분** | **설명** |
| 데이터 분포 확인 | * 예측할 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 |
| 머신러닝 모델 추천 | * 머신러닝 모델 학습 시작 * 머신러닝 모델 추천 결과 표 확인 * 머신러닝 모델 추천 결과 시각화 |

* 1. **테스트 데이터**
     1. 분류 모델
        1. Test\_01: mixing\_machine(교반구동장치)
        2. Test\_02: water\_potability
     2. 이상탐지 모델
        1. Test\_01: CNC Miling machine
        2. Test\_02: water\_potability
  2. **시스템 구성**

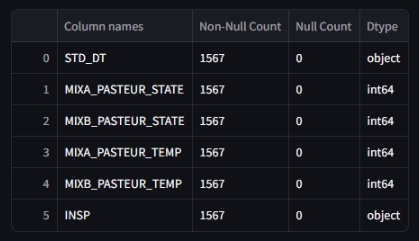


* + 1. Streamlit으로 개발한 클라이언트에서 FastAPI로 개발한 서버로 머신러닝 모델 학습 request.
    2. FastAPI로 개발한 서버에서 Streamlit으로 개발한 클라이언트로 머신러닝 모델 학습 결과 response.

1. **데이터 분포 학인 및 시각화**
   1. **Target Data 분포 확인**

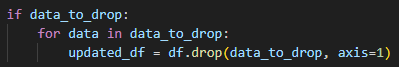


* + 1. 머신러닝 모델 학습 시 필요한 데이터 처리를 위한 상태 진단
  1. **데이터 전처리가 필요한 데이터 Type 확인**

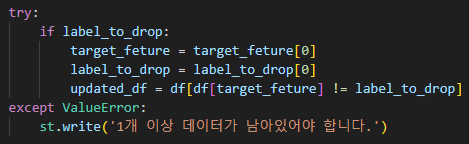


* + 1. 데이터 전처리가 필요한 데이터 Type 및 Null 값 분포 확인

1. **데이터 전처리**
   1. 불필요한 Columns 제거

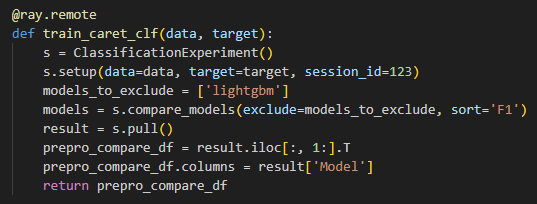


* + 1. Pandas를 사용하여 선택한 Column 제거
  1. 불필요한 Target Data 제거

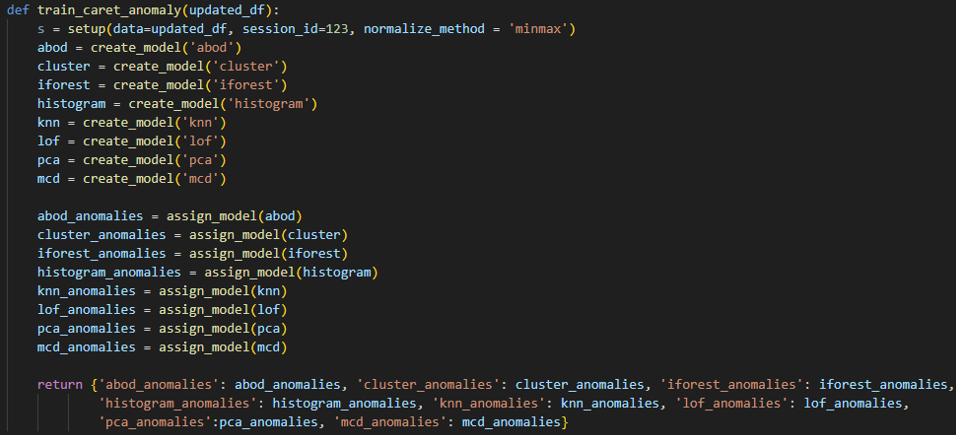


* + 1. Pandas를 사용하여 제거할 Label 제외하고 DataFrame 업데이트

1. **AutoML을 활용한 머신러닝 모델**
   1. **지도학습 머신러닝 분류 모델**

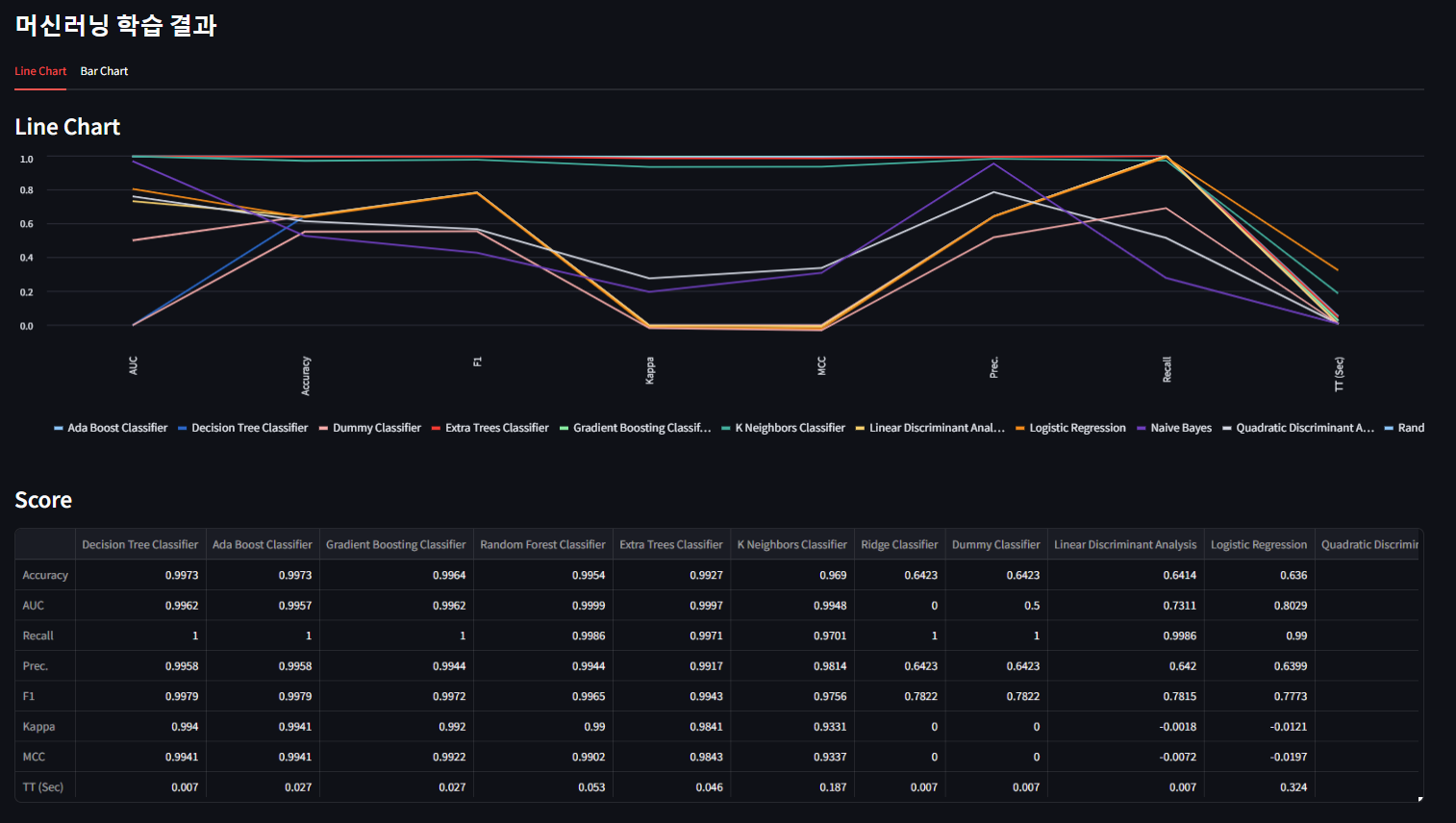


* + 1. F1-score를 기준으로 모델 검증
    2. AutoML 라이브러리 pycaret을 사용하여 약 16개 머신러닝 모델 사용.
    3. 병렬, 분산 처리 라이브러리 ray를 사용하여 AI 모델 학습 속도 개선.
    4. Lightgbm은 지나치게 느린 속도로 제외
  1. **비지도학습 머신러닝 이상탐지 모델**



* + 1. AutoML 라이브러리 pycaret에 있는 비지도학습 기반 이상탐지 모델 8개 학습.
    2. 병렬, 분산처리 라이브러리 ray를 사용한 것과 사용하지 않은 것에 속도 차이가 없어서 사용하지 않음.

1. **머신러닝 모델 추천**
   1. **지도학습 기반 분류 모델 결과 확인**

****

* + 1. 분류모델 학습 결과 Line Chart 시각화
    2. DataFrame에서 지표별 학습 결과 확인
    3. DataFrame 왼쪽부터 score가 높은 순서로 머신러닝 모델 추천.
  1. **비지도학습 기반 이상탐지 모델 결과 확인**



* + 1. 정상 데이터와 이상 데이터에 대한 anomaly score, counts 확인.
    2. 각 이상탐지 모델의 평균 score가 낮은 순서로 추천.
    3. 평균 score가 높을수록 이상치가 높음.



* + 1. 이상탐지 모델의 평균 이상치 score가 낮은 순서로 시각화 정렬