**SafetyPredictor**

**학번: 2218079**

**이름: 이은아**

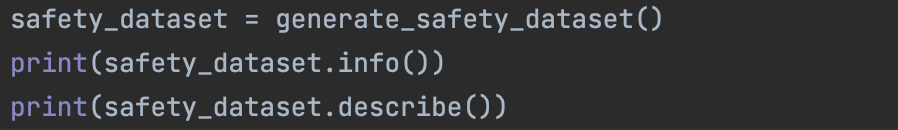
**Github address:** [**https://github.com/2218079LEEEUNA/-**](https://github.com/2218079LEEEUNA/-)

**e-mail : euna6467@naver.com**

1. **안전 관련 머신러닝 모델 개발의 목적** 
   1. 학습 모델 활용 대상: 화학물질을 처리하거나 다루는 사업장 내에서 일하는 모든 작업자들
   2. 화학물질 노출량과 안전 장비 사용 여부를 독립변수로 결정하고 그에 따라 안전사고 여부를 종속변수로 정했다. 만약, 안전사고 여부가 1이 나오면 안전사고가 발생한 것이고 0이면 안전사고가 발생하지 않음을 의미합니다.
   3. 개발의 의의: 화학물질 노출량과 안전 장비 사용 여부와 같은 입력 변수를 기반으로 안전사고 발생 여부를 예측하여 작업자들의 안전을 지키기 위한 목적이다. 이를 개발함으로써 안전 공학 분야에서 사고 발생 가능성을 사전에 예측하여 예방 조치를 취하거나 작업 환경을 개선하는 데 도움을 줄 수 있다.
2. **안전 관련 머신러닝 모델의 네이밍의 의미**

SafetyPredictor이라는 머신러닝 모델 이름은 모델에 대한 안전 관련된 정보를 다루고 안전사고 여부를 예측하기 때문에 SafetyPredictor이라고 이름을 붙였다.

1. **개발 계획**
   1. 데이터에 대한 요약 정리 및 시각화



이 코드를 통해 데이터셋에 대한 기본적인 정보와 수치형 특성의 기술 통계량을 출력한다.

* + - 1. Chemical\_Exposure(화학물질 노출량) : 가상의 데이터로, 0에서 100까지의 균일한 분포를 갖는 실수 값. 화학물질에 노출된 정도를 나타낸다.
      2. Safety\_Equipment\_Used(안전장비사용여부) : 가상의 데이터로, 0 또는 1의 이진값이다. 0은 안전 장비의 미사용을, 1은 안전 장비 사용을 나타낸다.
      3. Safety\_Incident(안전사고여부) : 가상의 데이터로, 0또는 1의 이진 값이다. 안전사고가 발생한 경우 1, 발생하지 않은 경우 0을 나타낸다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

이 코드는 각 특성 간의 상관 관계를 시각화하고, Chemical\_Exposure의 분포를 히스토그램으로, Safety\_Equipment\_Used 와 Safety\_Incident간의 관계를 막대 그래프로 나타낸다.

* 1. 데이터 전처리 계획

|  |  |
| --- | --- |
| 데이터 생성 | Generate\_safety\_dataset함수를 사용하여 화학물질 노출량, 안전 장비 사용 여부, 안전사고 여부 등의 속성을 가진 안전 관련 데이터셋을 생성한다. 이는 머신러닝 모델에 대한 학습 및 평가를 위한 가상의 데이터를 만드는 데 사용된다. |
| 데이터 준비 | 생성된 데이터셋에서 독립변수(x)와 종속 변수(y)를 분리한다. 독립 변수는 화학물질 노출량과 안전 장비 사용 여부를 포함하고, 종속변수는 안전사고 여부를 나타낸다. |
| 데이터 분할 | Train\_test\_split함수를 사용하여 전체 데이터를 훈련세트와 테스트 세트로 나눈다. 이를 통해 모델을 훈련하고 평가하는 데 사용한다. 훈련 세트는 모델 학습에 사용하고 테스트세트는 모델 평가에 사용한다. |
| 모델 학습 | Logistic regression모델을 선택하고, 훈련 세트를 사용하여 모델을 학습시킨다. Logistic Regression은 이진 분류 문제에 적합한 알고리즘 중 하나로, 확률을 기반으로 사고가 발생할 확률을 예측한다. |
| 모델평가 | 학습된 모델을 사용하여 훈련 세트와 테스트 세트에 대한 예측을 수행하고, 이를 사용하여 정확도 및 분류 보고서 등의 성능 지표를 계산한다. |
| 시각화 | Matplotlib를 사용하여 훈련 데이터와 테스트 데이터에 대한 예측 결과를 시각화한다. 이를 통해 모델의 학습 결과를 살펴볼 수 있다. |

* 1. 어떠한 머신러닝 모델을 사용할 것인지 (해당 머신러닝 모델의 이론 추가)

로지스틱 회귀(Logistic Regression)모델을 사용했다.

* + - 1. 로지스틱 회귀는 주로 이진 분류 문제에 사용되는 통계 기반의 머신러닝 알고리즘 중 하나이다.
      2. 명칭상 “회귀”이지만 실제로는 분류 문제에 사용한다.
      3. 주로 선형 결합을 로지스틱 함수에 통과시켜 확률 값을 계산하고, 그 확률 값을 기반으로 샘플을 클래스에 할당한다.

로지스틱 함수(시그모이드 함수)

수학적 표현 :

(여기서 z는 선형 결합(가중치와 특성의 합)을 나타내며, e는 자연 상수(2.71828…)이다.)

* 1. 머신러닝 모델 예측 결과가 어떠할 지

훈련된 로지스틱 회귀 모델은 새로운 입력 특성에 대해 확률을 출력한다. 일반적으로 0.5를 기준으로 이진 분류를 수행하며, 확률값이 0.5보다 크면 양성 클래스로 1이고, 그렇지 않으면 음성 클래스 0으로 예측한다.

* 1. 사용할 성능 지표

코드에서 주로 사용한 성능지표는 정확도, 분류 보고서, 혼동 행렬이다.

* + - 1. 정확도
         * 전체 예측 중 올바르게 예측한 비율
         * 공식 : (TP+TN)/(TP+TN+FP+FN)
         * 코드에서는 ‘accuracy\_score’함수를 사용하여 계산
      2. 분류 보고서
         * 각 클래스에 대한 정밀도, 재현율, F1점수를 제공
         * 정밀도 : 양성으로 예측한 것 중 실제 양성의 비율, 공식 : TP/(TP+FP)
         * 재현율 : 실제 양성 중에서 양성으로 예측한 비율, 공식 : TP/(TP+FN)
         * F1점수 : 정밀도와 재현율의 조화 평균. 공식 : 2x(정밀도x재현율)/(정밀도+재현율)
         * 코드에서는 classification\_report 함수 사용
      3. 혼동 행렬
         * 각 클래스에 대한 실제 레이블과 모델의 예측 결과에 따라 예측 결과를 나타낸다.
         * TP(TRUE Positive) : 실제 양성. 모델도 양성 예측
         * TN(TRUE Negative) : 실제 음성. 모델도 음성 예측
         * FP(FALSE Positive) : 실제 음성. 모델 양성 예측
         * FN(FALSE Negative) : 실제 양성. 모델 음성 예측
         * 코드에서는 confusion\_matrix함수 사용하여 계산
  1. 성능 검증 방법 계획 등

|  |  |
| --- | --- |
| 데이터 분할 | 데이터를 훈련세트, 검증세트, 테스트 세트로 나눈다. |
| 모델 훈련 | 로지스틱 회귀 모델을 사용하여 안전사고 예측 모델을 훈련 |
| 모델 검증 및 평가 | 검증 세트를 사용 하여 모델의 성능을 평가 한다. 정확도, 정밀도, 재현율, F1점수 등을 계산한다. |
| 하이퍼파라미터 튜닝 | 텍스트, 스크린샷, 폰트이(가) 표시된 사진  자동 생성된 설명  모델의 성능을 향상시키기 위해 하이퍼파라미터를 조정한다.  교차 검증을 사용하여 다양한 하이퍼파라미터 조합을 평가한다. |
| 최종 성능 평가 | 텍스트, 폰트, 스크린샷이(가) 표시된 사진  자동 생성된 설명  테스트 세트를 사용하여 최종 모델의 성능을 평가한다.  정확도, 정밀도, 재현율, F1점수 등을 계산한다. |
| 모델 해석 및 향상 | 모델이 어떻게 예측을 수행했는지 해석하고, 안전사고에 영향을 미치는 주요 특성을 식별한다. 그리고 결과를 관리자나 안전 전문가에게 전달하여 안전 조치의 개선에 활용한다. |
| 지속적인 모니터링 및 개선 | 모델의 성능을 지속적으로 모니터링하고 필요한 경우 개선 작업을 수행한다. 새로운 데이터를 수집하고 모델을 업데이트 하여 안전사고 예측 능력을 최적화한다. |

1. **개발 과정**

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

필요한 머신러닝 모델과 데이터 처리 및 평가를 위한 라이브러리들을 임포트 한다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

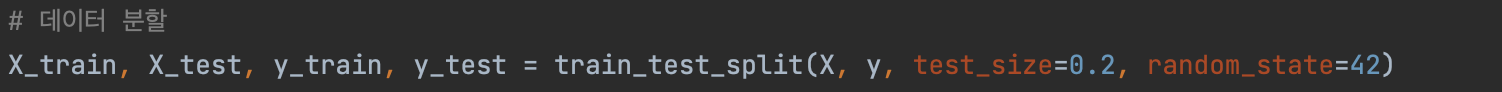
자동 생성된 설명

안전 관련 데이터 셋을 생성하는 함수를 정의한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

생성된 데이터셋을 사용하여 독립 변수(X)와 종속 변수(Y)를 준비한다.



데이터를 훈련 세트와 테스트 세트로 분할한다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

로지스틱 회귀 모델을 선택하고, 훈련 데이터를 사용하여 모델을 훈련시킨다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

훈련된 모델을 사용하여 훈련과 테스트에 대한 예측을 수행한다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

정확도를 사용하여 훈련과 테스트에 대한 모델의 성능을 평가한다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

Classification\_report 와 confusion\_matrix를 사용하여 분류보고서와 혼동 행렬을 출력한다.

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

Matplotlib를 사용하여 훈련과 테스트에 대한 예측 결과를 시각화 한다.

스크린샷, 라인, 그래프, 직사각형이(가) 표시된 사진

자동 생성된 설명

1. **개발 후기**

**요번 과제를 수행하면서 정말 몇 번을 주제를 바꾸고 처음부터 다시 수행해보기를 반복했다. 사실 지금도 내가 한 과제가 맞는 답일지 헷갈릴 정도로 과제를 하는데 있어 어려움이 많았고 이해하기가 힘들었다. 하지만 이런 과정을 수행하면서 내가 지금 해낸 과제가 올바른 정답이 아니더라도 분명 얻은 점은 많다고 생각한다. 과제에 투자한 시간이 정말 많기 때문에 포기하고 싶지 않아서 유튜브 등을 찾아보면서 과제를 하였는데 그 과정에서 공부가 많이 되었다. 사실 교수님과 요번 컴퓨터 프로그래밍을 배우면서 프로그래밍이 너무 재밌고 흥미로워서 Ai공학과로의 전과도 생각했었다. 하지만 내가 지금까지 배운 건 정말 빙산의 일각도 안되는구나 라는 것을 요번 과제를 하면서 알게 되었다. 그래도 요번 학기 동안 교수님과 프로그래밍에 대해 배우면서 새로운 길들과 프로그램이 개발되는 과정을 조금이나마 알게 된 것 같아서 너무 유익한 시간이었다.**