# Industrial-Safety Prediction Model

**이름: 안소연**

**학번: 2018030**

**Github: https://github.com/soyeon-ann/Industrial-Safety-Prediction-Model.**

## 1. 안전 관련 머신러닝 모델 개발 관련 요약

본 프로젝트는 산업 기계의 고장 가능성을 예측하는 머신러닝 모델을 개발하였다. 기계의 성능 데이터를 기반으로 고장 발생 가능성을 예측하기 위해 랜덤 포레스트(Random Forest) 모델을 활용하였다. 이를 통해 작업 현장의 안전성을 향상시키고 예기치 못한 사고를 예방하고자 하였다.

## 2. 개발 목적

### a. Industrial Safet Prediction Model 활용대상

i. 제조업
   1) 생산라인
   2) 설비 유지보수
   3) 품질 관리

ii. 물류
   1) 자동화 창고 시스템
   2) 배송차량

iii. 교통시스템
   1) 철도 및 전철 시스템
   2) 항공기 정비

### b. 개발의 의의

#### i. 사고 예방 및 안전성 향상

안전 머신러닝 모델은 산업 기계의 고장을 사전에 예측하여 예기치 않은 사고를 방지한다. 이를 통해 작업 현장의 안전성을 높이고, 기계 고장으로 인한 안전사고를 줄이며, 기업의 경제적 손실을 최소화할 수 있다.

### ii. 유지보수 및 생산 효율성 최적화

모델을 활용해 기계 고장을 정확히 예측함으로써, 적시에 유지보수를 수행하고 불필요한 유지보수 비용을 절감할 수 있다. 또한, 고장 예측을 통해 생산 중단을 방지하고 생산 효율성을 높이는 효과가 기대된다.

## c. 데이터의 독립 변수와 종속 변수

i. 독립변수:

1) Air temperature [K]: 공기 온도 [K]
2) Process temperature [K]: 공정 온도 [K]
3) Rotational speed [rpm]: 회전 속도 [rpm]
4) Torque [Nm]: 토크 [Nm]
5) Tool wear [min]: 공구의 마모 정도 [분]

ii. 종속변수:

1) Machine failure (0: No failure, 1: Failure): 기계 고장 (0: 고장 없음, 1: 고장 발생)

# 3. 배경지식

## a. 데이터 관련 사회 문제

기계 고장의 예측 실패는 산업 현장에서 심각한 안전사고를 초래할 수 있다. 예를 들어, [1]2015 년 한화케미칼 울산 2 공장에서 발생한 폭발사고는 폐수처리 기계의 고장을 예측하지 못해 발생한 사고이다. 이 사고로 6 명이 사망하고 여러 명이 부상을 입었으며, 사고의 원인은 기계 고장 예측 시스템의 부재로 분석되었다. 이러한 사고는 단순한 인명 피해에 그치지 않고, 생산 중단과 경제적 손실을 초래하는 등 기업에 큰 영향을 미친다.

이러한 문제를 해결하기 위해, 머신러닝 기반의 고장 예측 시스템이 중요한 역할을 할 수 있다. 기계의 상태를 실시간으로 모니터링하고 데이터를 분석함으로써 고장을 사전에 예측하고, 사고를 예방할 수 있다. 이를 통해 기업은 사고를 미연에 방지하고, 안전성을 높일 수 있다.

[1] . <한화케미칼 사고, 기계 고장 예측 실패> , https://www.safety1st.news/news/articleView.html?idxno=5719&utm_source=chatgpt.com.

## 4. 개발 내용

### a. 데이터에 대한 구체적 설명 및 시각화

i. 데이터 개수

1) 행 개수: 10,000
2) 열 개수: 14

ii. 데이터 속성

1) 'Air temperature [K]' (공기 온도 [K])
2) 'Process temperature [K]' (공정 온도 [K])
3) 'Rotational speed [rpm]' (회전 속도 [rpm])
4) 'Torque [Nm]' (토크 [Nm])
5) 'Tool wear [min]' (툴의 마모 정도 [분])
6) 'Machine failure' (기계 고장 여부, 0: 고장 없음, 1: 고장 발생)

iii. 데이터 간 상관관계 분석

Air temperature [K]와 Process temperature [K]는 매우 강한 상관관계(0.876)를 보인다. 또한, 'Machine failure'와 'HDF'는 중간 수준의 상관관계(0.576)를 보인다.

### b. 머신러닝 모델 선정 이유

i. 랜덤 포레스트 모델 선정 이유

랜덤 포레스트는 여러 개의 결정 트리를 사용하여 예측을 수행하고, 이들 트리의 예측을 종합하여 더 정확한 예측 결과를 도출한다. 특히, 데이터셋에 포함된 다양한 변수들 간의 복잡한 관계를 잘 처리할 수 있어, 고장 예측 문제에 적합한 모델로 판단하였다.

ii. 성능 비교를 위한 머신러닝 모델 선정 이유

랜덤 포레스트는 여러 트리를 활용해 예측을 수행하며, 각 트리의 약점을 보완하여 예측 성능을 높인다. 다른 머신러닝 모델들에 비해 과적합을 방지할 수 있는 장점이 있어, 안정적이고 신뢰성 있는 예측 결과를 제공한다. 성능 비교를 통해 랜덤 포레스트가 가장 우수한 성능을 보였으며, 고장 예측에 적합한 모델로 선택되었다.

## c. 사용할 성능 지표

### i. 성능 지표 설명

1) 정확도 (Accuracy):
전체 예측에서 올바르게 예측한 비율을 나타내며, 모델이 고장 예측을 얼마나 정확하게 수행하는지 평가하는 지표이다.

2) 정밀도 (Precision):
양성으로 예측한 것 중 실제로 양성인 비율을 나타낸다. 즉, 모델이 예측한 양성 중에서 실제로 맞은 비율이다.

3) 재현율 (Recall):
실제 양성 중에서 모델이 양성으로 정확하게 예측한 비율을 나타낸다. 즉, 실제 양성인 데이터가 얼마나 잘 예측되었는지를 평가한다.

4) F1 Score:
Precision 과 Recall 사이의 균형을 맞추는 데 유용하며, 두 지표를 종합적으로 고려할 때 유용한 지표이다.

5) 혼동 행렬 (Confusion Matrix):
실제값과 예측값을 비교하여, 모델의 예측 성능을 시각적으로 확인할 수 있는 지표이다.

### ii. 성능 지표 선정 이유

정확도는 모델이 전체 데이터에서 얼마나 정확하게 예측했는지를 나타내며, 모델의 기본 성능을 평가하는 데 유용하다. 정밀도는 모델이 예측한 긍정 클래스 중 실제로 맞춘 비율을 측정하며, 잘못된 경고를 줄이는 데 중요한 지표이다. 재현율은 실제 긍정 클래스 중 모델이 올바르게 예측한 비율을 나타내며, 고장을 놓치지 않기 위해 중요한 평가 지표이다.
F1 점수는 정밀도와 재현율의 균형을 고려한 지표로, 두 지표의 균형을 맞추는 데 유용하다.

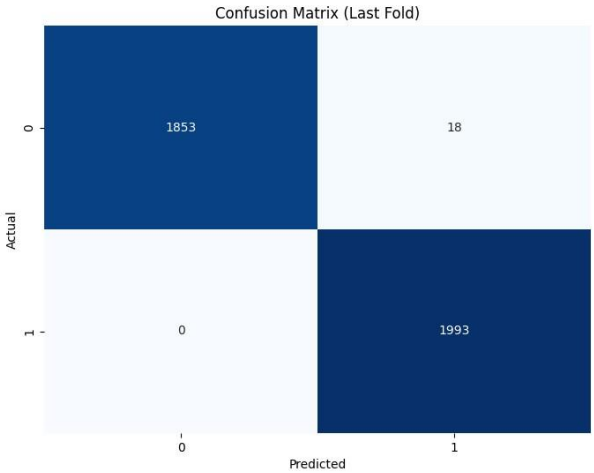
## 5. 개발 결과

### a. 성능 지표에 따른 머신러닝 모델 성능 평가

본 프로젝트에서는 Accuracy, Precision, Recall, F1 Score, Confusion Matrix 와 같은 성능 지표를 사용하여 모델의 성능을 평가하였다. 다음은 최종 모델의 평가 결과이다.

i. Accuracy: 99.72%
ii. Precision: 0.99
iii. Recall: 1.00
iv. F1 Score: 1.00


v. Confusion Matrix



## b. 머신러닝 모델의 성능 결과에 대한 해석

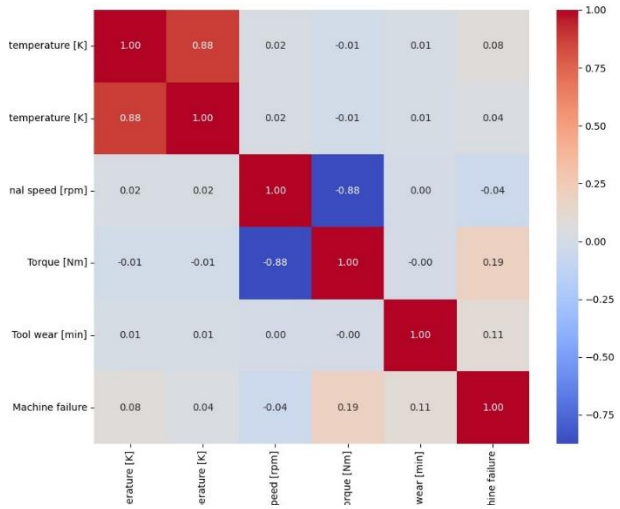
i. K-Fold Cross Validation Results

모델의 성능을 평가하기 위해 5-Fold Cross Validation 을 수행하였다. 각 Fold 에서의 성능 평가 결과를 평균으로 계산한 주요 성능 지표는 다음과 같다.

- **Average Accuracy**: 1.00
  (모델이 전체 데이터에서 정확히 예측한 비율)
- **Average Precision**: 0.99
  (모델이 예측한 긍정 클래스 중 실제로 긍정 클래스인 데이터의 비율)
- **Average Recall**: 1.00
  (실제 긍정 클래스를 모델이 놓치지 않고 예측한 비율)
- **Average F1 Score**: 1.00
  (Precision 과 Recall 의 조화 평균으로, 모델의 전반적인 예측 성능을 종합적으로 평가)

따라서 K-Fold Cross Validation 결과에서 모든 지표에서 거의 완벽에 가까운 성능을 보였다. 특히 Accuracy 와 Recall 이 1.00 으로 측정되어, 데이터의 모든 클래스에 대해 우수한 예측 성능을 나타낸다. 그리고 Precision 이 0.99 로 약간 낮은 것은 일부 False Positive 가 존재할 가능성을 나타낼 수 있다.

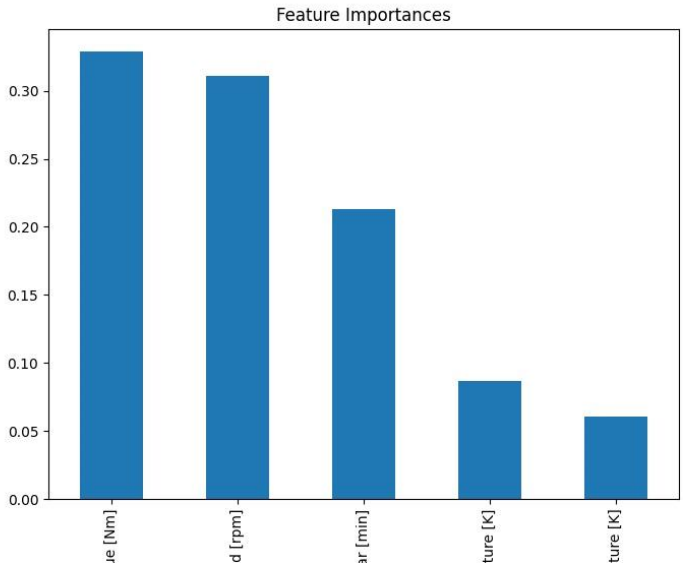
ii. 상관 관계 히트맵



'Air temperature [K]' (공기 온도 [K])와 'Process temperature [K]' (공정 온도 [K]) 와의 관계가 가장 높은 양의 상관관계를 나타내고 있다. 이는 공기 온도가 공정 온도의 변화에 중요한 영향을 미친다는 것을 알 수 있다. 결과적으로 공정 온도가 산업 기계의 작동 상태가 열적 안정성에 영향을 줄 수 있음을 의미한다.

즉, 기계 내부 부품의 과열, 냉각 효율 저하, 또는 재료 열화와 같은 요인으로 이어질 가능성이 있으며, 이는 궁극적으로 기계 고장의 주요 원인으로 작용할 수 있다. 공기 온도와 공정 온도 간의 관계를 이해하고 이를 모니터링 하는 것이 본 모델의 정확성을 높이는데 기여할 수 있다.

iii. Feature importances




'Rotational speed', 'Torque', 'Tool wear' 변수가 기계 고장 가능성을 예측하는 데 있어 매우 중요한 역할을 하는 것으로 나타난다. 'Rotational speed'와 'Torque'는 기계의 실시간 작동 상태와 과부하를 나타내며, 'Tool wear'는 기계 유지보수 주기를 결정짓는 중요한 요인으로 고장을 예방하고 생산성을 유지하는 데 기여한다.
이 변수들 간의 데이터를 지속적으로 분석하면 고장 위험이 높은 시점을 정확히 예측하고 적절한 예방 조치를 통해 유지보수 효율성을 극대화할 수 있다.

## 6. 결론

### a. 머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명

본 프로젝트에서는 산업 기계의 고장 가능성을 예측하기 위해 RandomForestClassifier 를 사용한 머신러닝 모델을 개발하였다. 데이터는 전처리를 통해 결측치를 제거하고, 클래스 불균형 문제를 해결하기 위해 RandomOverSampler 를 활용하였다. KFold 교차 검증 결과 평균 정확도(1.00), 정밀도(0.99), 재현율(1.00), F1 스코어(1.00)로 매우 우수한 성능을 달성하였다. 또한, Feature Importance 분석 결과, '회전 속도', '토크', '툴의 마모 정도'가 가장 중요한 변수로 나타났다.

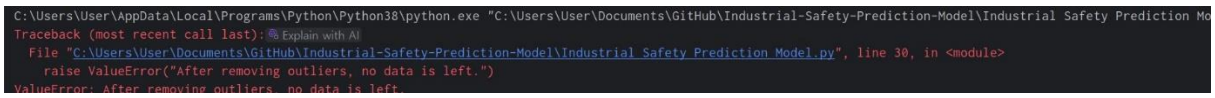
### b. 개발 의의

본 모델은 산업 현장에서 기계 고장을 사전에 예측하여 예방 조치를 취할 수 있는 도구로 활용될 수 있다. 특히, 중요한 변수로 나타난 '회전 속도', '토크', '툴의 마모 정도'는 기계 상태 모니터링 및 유지보수 전략 수립에 중요한 데이터를 제공할 수 있다. 이를 통해 산업 안전성을 강화하고 생산성을 향상시키는 데 기여할 것으로 기대된다.

### c. 머신러닝 모델의 한계

모델의 성능은 데이터의 품질과 양에 크게 의존하며, 본 프로젝트는 제한된 데이터 셋을 사용하여 학습 및 평가를 진행하였다. 또한, 모든 KFold 교차 검증 결과가 100%에 가까운 정확도를 보였지만, 이는 데이터가 모델에 과적합되었거나 데이터의 분포가 실제 산업 환경과 다를 가능성을 배제할 수 없다. 따라서 더 다양한 데이터와 실제 산업 데이터를 추가로 수집하여 모델을 검증할 필요가 있다.

### d. 머신 러닝 개발을 하며 발생하였던 오류




결측치 제거 및 이상치 처리 과정에서, 데이터가 거의 남지 않게 되어 학습에 사용할 수 없는 치명적인 오류가 발생하였다. 이는 원래 데이터에서 결측치와 이상치가 상당한 비율을 차지하고 있었기 때문이다. 결측치와 이상치의 존재는 모델 학습의 신뢰성을 저하시킬 수 있으며, 이러한 문제를 해결하지 않으면 모델이 잘못된 패턴을 학습하거나 예측 성능이 떨어질 수 있기 때문에 전처리 과정은 필수적이었다. 그러나 전처리 후에도 데이터 부족 문제는 여전히 해결되지 않았고,

학습에 사용할 수 있는 데이터가 극히 제한적이게 되었다. 이러한 문제를 극복하기 위해 RandomOverSampler 를 사용하여 클래스 불균형을 해결하고 데이터 세트를 보강하였다.
RandomOverSampler 와 같은 기법을 사용한 경우, 소수 클래스의 데이터가 중복 복제되면서 과적합(overfitting) 문제를 발생시킬 수 있다는 우려가 있었다.
이를 해결하기 위해 교차 검증(Cross-Validation)을 통해 모델의 일반화 능력을 테스트하고, SMOTE 나 언더샘플링 등 다양한 불균형 처리 기법을 시도하였다. 또한, 정확도 외에도 정밀도(precision), 재현율(recall), F1-score 와 같은 다양한 성능 지표를 활용하여 모델을 평가하였다. 이를 통해 데이터 재구성 후에도 모델의 신뢰성을 높일 수 있었다.

### e. 느낀 점

프로젝트를 시작할 때, 안전 관련 머신러닝이라는 주제를 정하면서 무엇을 해야 할지 고민이 정말 많았다. 안전이라는 주제가 워낙 포괄적이다 보니, 적합한 데이터셋을 찾고 이를 주제로 삼는 데 시간이 많이 걸렸다.

개발을 진행하며 머신러닝 모델 개발은 단순히 데이터를 입력하고 결과를 얻는 과정이 아니라는 것을 깨달았다. 특히, 데이터 전처리와 클래스 불균형 문제를 해결하는 과정에서 모델 성능에 얼마나 많은 영향을 미치는지를 체감할 수 있었다. 그리고 요즘처럼 취업시장이 어려운 상황에서, 이런 프로젝트 경험은 내게 많은 생각을 하게 해줬다. '안전공학과를 전공했고, 여러 관련 자격증이 있으니 파이썬에 대한 내용을 좀 더 열심히 학습하여 이를 융합시켜 무언가를 개발하는 것도 미래에 도움이 될 수 있겠구나' 하는 생각이 들었다.

특히, 인공지능이 많이 발전하면서 기본적인 개발은 인공지능의 도움을 받을 수 있지만, 그 기반을 다지는 과정과 어떠한 명령 프롬프트를 사용하며 진행할 지 등 이를 어떻게 활용할지를 고민하는 건 여전히 내 몫이라는 생각이 들었다.
마지막으로 이번 프로젝트는 단순히 기술을 익히는 것을 넘어, 내가 무엇을 잘할 수 있고, 앞으로 어떤 방향으로 나아가야 할지를 고민하는 계기가 되었다. 더 다양한 데이터를 다뤄보고, 실제 문제를 해결할 수 있는 모델을 만드는 과정에서 부족함을 느끼기도 했지만, 동시에 내가 가진 역량을 더욱 발전시켜 나아가야겠다고 다짐할 수 있었다.