# LPG 누출 탐지 머신 러닝 모델 개발

**이름: 하유승**

**학번: 2118040**

**Github:**

## 1. 안전 관련 머신 러닝 모델 개발 관련 요약

본 프로젝트의 목적은 LPG 누출을 탐지하기 위한 머신 러닝 모델을 개발하는 것이다. 이를 통해 산업 현장에서의 안전성을 높이고, 잠재적인 사고를 예방하고자 한다.

## 2. 개발 목적

a. 머신 러닝 모델 활용 대상: LPG 를 다루는 다양한 분야의 산업 현장
개발의 의의: 다양한 화학물질을 다루는 산업 현장에서 LPG 가스의 누출 여부를 판단하여 사전에 화재를 예방할 수 있다. IoT 센서를 통해 실시간으로 탐지되는 다양한 화학물질 속에서 LPG 가스 누출 여부를 사전에 파악하여 사고를 미연에 방지하고 빠른 대응이 가능해진다.

b. 데이터에 대한 구체적 설명 및 시각화:
1001 개의 데이터 및 9 개의 속성을 가진 "7.lpg_leakage.csv"는 다양한 화학물질을 다루는 산업 현장에서 IoT 센서를 통해 탐지되는 다양한 화학물질 속에서 LPG 가스의 누출 여부를 판단하여 사전에 화재 예방을 돕는다. 현장에서 평소에 존재하는 LPG 의 양인지 혹은 가스관 파열로 인한 가스 과다 누출인지 여부를 판단하여 대형 사고를 막을 수 있다.

c. 데이터의 어떠한 독립 변수를 사용하여 어떠한 종속 변수를 예측하는지
- 독립변수: Alcohol, CH4, CO, H2, LPG, Propane, Smoke, Temp
- 종속변수: LPG_Leakage

## 3. 배경지식

a. 데이터 관련 사회 문제 설명

LPG 는 인화성이 강한 가스로, 누출 시 화재나 폭발 사고로 이어질 수 있다. 특히 산업 현장에서는 대량의 LPG 를 사용하기 때문에 누출 사고가 발생할 경우 큰 피해를 초래할 수 있다.

b. 머신러닝 모델 관련 설명 등

머신러닝 모델은 대량의 센서 데이터를 분석하여 패턴을 인식하고, LPG 누출 여부를 예측하여 실시간 모니터링과 빠른 대응이 가능하다.

본 프로젝트에서는 로지스틱 회귀, 랜덤 포레스트, SVM 모델을 사용하여 예측 모델을 개발했다. 각 모델의 특성과 장단점을 비교하여 최적의 모델을 선정했다. 데이터의 품질을 높이기 위해 결측값 처리, 이상치 제거, 데이터 정규화 등의 전처리 과정을 거쳤다.

## 4. 개발 내용

a. 데이터에 대한 구체적 설명 및 시각화

데이터 개수: 1001 개

데이터 속성: Alcohol, CH4, CO, H2, LPG, Propane, Smoke, Temp, LPG_Leakage

b. 데이터에 대한 설명 이후, 어떤 것을 예측하고자 하는지 구체적으로 설명

8 가지의 독립 변수들 중 LPG 누출과 직접적으로 관련 있는 변수들을 찾아내고 LPG 누출이 일어날 것인가를 예측한다.

c. 머신러닝 모델 선정 이유

LPG 는 산업 현장에서 매우 많이 쓰이는 화학 물질들 중 하나이다. 그러면서도 매우 사고가 많이 일어나는 물질들 중 하나이기도 하다. 따라서 산업 현장의 가스 누출을 파악하는 것이 사고 예방 건수를 유의미하게 낮출 수 있다고 보았다.

d. 사용할 성능 지표

**로지스틱 회귀**는 모델의 출력이 각 독립 변수의 가중치로 표현되기 때문에, 변수들이 결과에 미치는 영향을 쉽게 해석할 수 있으며 누출 여부와 같이 이진 분류에 적합하다.

**랜덤 포레스트**는 다양한 변수와 복잡한 상호작용을 고려하여 높은 예측 성능을 제공하며, 여러 개의 결정 트리를 앙상블하여 각 변수의 중요도를 계산해낸 뒤 주요한 예측 변수를 구별해낼 수 있다.

**서포트 벡터 머신**은 수많은 변수가 존재하는 고차원 데이터에서도 효과적으로 작동하며, 결정 경계와 데이터 포인트 간의 마진을 최대화하여 분류 성능을 높인 모델이다.

## 5. 개발 결과

### 성능 지표에 따른 머신러닝 모델 성능 평가

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Logistic Regression | 0.85 | 0.820513 | 0.8 | 0.810127 |
| Random Forest Classifier | 1.0 | 1.0 | 1.0 | 1.0 |
| Support Vector Machine Classifier | 0.93 | 0.923077 | 0.9 | 0.911392 |

### K-Fold 교차 검증 결과

| Model | Cross-Validation Scores | Mean Accuracy |
|---|---|---|
| Logistic Regression | [0.855, 0.835, 0.88, 0.86, 0.81] | 0.848 |
| Random Forest Classifier | [1.0, 0.995, 0.995, 0.995, 1.0] | 0.997 |
| Support Vector Machine Classifier | [0.87, 0.865, 0.89, 0.86, 0.84] | 0.865 |

a. 머신러닝 모델의 성능 결과에 대한 해석

로지스틱 회귀 모델은 해석이 용이하고 계산 속도가 빠르며, 이진 분류 문제에 적합하다. 그러나 다른 모델에 비해 정확도와 F1 점수가 낮아, 복잡한 데이터에서는 성능이 떨어질 수 있다.

랜덤 포레스트 모델은 높은 예측 성능을 제공하며, 변수 중요도를 파악할 수 있습니다. 과적합을 방지하고 다양한 데이터에서 유연하게 작동한다. 그러나 모든 성능 지표가 완벽한 것은 과적합의 가능성을 시사할 수 있으므로, 교차 검증을 통해 일반화 성능을 확인하는 것이 중요하다.

SVM 모델은 고차원 데이터에서도 효과적으로 작동하며, 마진을 최대화하여 분류 성능을 향상시킬 수 있다. 다양한 커널 함수를 사용하여 비선형 데이터를 처리할 수 있다. 그러나 계산 비용이 높고, 대규모 데이터셋에서는 학습 시간이 오래 걸릴 수 있다.

## 6. 결론

**랜덤 포레스트 모델**은 높은 예측 성능과 변수 중요도 파악이 가능하며, 과적합을 최대한 방지할 수 있어 셋 중 최적의 모델로 선정하였다.

이렇게 여러 가지 머신러닝 모델을 활용하여 LPG 누출 탐지에 최적화된 모델을 찾는 과정을 거치면서 느낀 점은 상황에 따라 모델을 적절히 해석하여야 한다는 것이다.

앞으로 인공지능 시대에 대비하기 위해서는, 수많은 데이터들을 정리하고 취합하는 습관을 들이는 것이 살아남는 길임을 깨닫게 되었다. 그래서 본인이 지금껏 활동했던 일들을 생각날 때마다 적어 놓았다가, 데이터가 모이면 앞으로의 진로에 관하여 판단해보는 시간을 가져볼 계획이다.