LPG 누출 탐지 머신러닝 모델 개발

이름: 김기태

학번: 2118301

Github: gitae1234

목차

1. 요약

목표

- LPG 누출 사고를 예방하기 위한 머신러닝 기반 탐지 모델 개발한다.
- 센서 데이터를 기반으로 누출 여부를 실시간으로 예측하여 안전성을 강화한다.

주요 내용

- 데이터 분석, 전처리 및 시각화를 수행
- 로지스틱 회귀, 랜덤 포레스트 모델 등을 활용해 성능 비교
- 최적의 모델을 선정하고, 모델 성능을 평가

2. 개발 목적

활용 대상

- 화학 공장, 가정용 LPG 설비, LPG 충전소 등

의의

- **사회적 가치 창출:** 사고를 방지 및 예방함에 따라 작업자의 안전을 강화함으로써 인명과 재산 피해를 줄인다.
- **기술 혁신:** 기존 감지 시스템에 머신러닝 기술을 도입하여 데이터 기반의 스마트한 안전관리 시스템을 제공한다.
- **산업 표준화 기여:** 효율적인 시설 운영 및 유지보수 비용 절감등의 기대효과로 신뢰성 있는 모델을 통해 관련 업계의 안전관리 표준을 높이는 데 기여한다.

독립 변수 및 종속 변수

- **독립 변수**: 센서 데이터 (Alcohol, CH4, CO, H2, LPG, Propane, Smoke, Temp).
- **종속 변수**: LPG_Leakage (누출 여부: 1-누출, 0-정상).

3. 배경지식

LPG(액화석유가스)는 산업 및 가정용 에너지원으로 널리 사용되지만, 누출 시 폭발 위험과 독성 가스 흡입 위험을 초래할 수 있다. 주요 문제는 누출을 조기에 탐지하지 못할 경우, 인적·물적 피해가 심각하다는 것이다. 기존 누출 탐지 기술은 주로 고정 센서를 사용하며, 이로 인해 누출 위치를 정확히 파악하거나 소규모 누출을 실시간으로 감지하는 데 한계가 있다.

머신러닝 모델은 다량의 센서 데이터를 분석하여 보다 빠르고 정확한 누출 탐지를 가능하게 하며, 안전성을 높이고 비용을 절감하는 데 기여할 수 있다. 특히, 센서 데이터는 다양한 물질(예: CH4, Propane, Smoke 등)의 농도를 포함하여 누출 여부를 결정하는 데 중요한 요소로 작용하기 때문에 데이터를 효율적으로 처리하고 누출 여부를 실시간으로 예측할 수 있는 머신러닝 모델이 필수적이다.

4. 개발 내용

a. 데이터에 대한 구체적 설명 및 시각화

i. 데이터 개수 및 데이터 속성

- 데이터는 총 1000 개의 샘플로 구성되어 있으며, 8 개의 독립 변수와 1 개의 종속 변수로 이루어져 있습니다.
- 독립 변수는 센서 데이터를 나타내며, 각각 Alcohol, CH4, CO, H2, LPG, Propane, Smoke, Temp 로 명명되었습니다.
- 종속 변수인 LPG_Leakage 는 누출 여부를 이진 값(0: 정상, 1: 누출)으로 나타냅니다.

ii. 데이터 간 상관관계

- 각 변수 간의 상관계수를 분석하여 주요 변수 간 관계를 확인했다.
- 특히, LPG 와 LPG_Leakage 간의 상관계수가 높게 나타나 예측 모델에서 중요한 역할을 할 것으로 판단된다.
- 상관계수를 분석한 결과, 독립 변수 중 Smoke 와 Propane 도 의미 있는 상관성을 보여 추가적인 분석이 요구되었다.

b. 데이터에 대한 설명 이후 예측 목표

iii. 예측 목표

- 센서 데이터를 기반으로 LPG 누출 여부를 정확히 예측하여 사고를 사전에 방지하는 데에 있다.
- False Negative(누출을 탐지하지 못하는 경우)를 최소화하는 것이다.

c. 머신러닝 모델 선정 이유

iv. 선정 이유

- **로지스틱 회귀**: 모델 구조가 단순하며 해석이 용이하여 초기 분석에 적합함으로 선정함.
- **랜덤 포레스트**: 비선형 데이터를 효과적으로 처리할 수 있고 높은 예측 성능을 제공한다. 특히, 변수 간 상호작용을 잘 포착하여 복잡한 데이터셋에도 강점을 보이기 때문에 선정함.
- **성능비교**: 다양한 머신러닝 모델을 적용해보며 최적의 성능을 제공하는 모델을 선정하기 위해 여러 모델을 비교하였으며, 단일 모델 성능뿐 아니라 교차 검증을 통해 모델의 안정성을 평가하였다.

d. 사용 성능 지표

v. 사용 지표

- 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1 Score), 혼동 행렬(Confusion Matrix) 등 다양한 지표를 사용하였다.
- 재현율(Recall)은 누출을 탐지하는 모델의 성능을 평가하는 데 가장 중요한 지표로 선정되었다.

vi. 선정 이유

- 정확도 외에도, 재현율과 정밀도를 함께 고려하여 안전 관리 측면에서 놓치기 쉬운 누출을 더 잘 탐지할 수 있도록 평가 지표를 설정하였다.
- F1 Score 는 정밀도와 재현율의 조화 평균으로, 모델의 전반적인 균형을 평가하는 데 사용되었다.

5. 개발 결과

a. 성능 지표에 따른 머신러닝 모델 성능 평가

- **로지스틱 회귀**:
  - 정확도: 약 88%
  - 주요 변수로는 LPG, CH4 가 높은 기여도를 나타냈다.
  - 혼동 행렬 분석 결과, 일부 False Negative 사례가 발생하였으나 False Positive 보다 상대적으로 적었음.
- **랜덤 포레스트**:
  - 정확도: 약 92%
  - 주요 변수로는 LPG, Propane, Smoke 가 높은 기여도를 나타냈다.
  - 결과적으로 False Negative 가 최소화되어 실질적인 누출 탐지 성능이 뛰어났다.
  - 랜덤 포레스트는 주요 변수 중요도를 산출하여 모델 해석 가능성을 높였다.

b. 머신러닝 모델의 성능 결과 해석

- 랜덤 포레스트 모델은 정확도뿐만 아니라 재현율에서도 높은 점수를 기록하여 안전 관리에 적합한 모델로 평가되었다.
- 로지스틱 회귀는 모델이 단순하지만, 일부 복잡한 비선형 데이터를 처리하는 데 한계를 보였다.
- 교차 검증 결과, 랜덤 포레스트 모델이 가장 일관된 성능을 보여 최적의 모델로 선정되었다.

6. 결론

a. 요약

- 머신러닝 기반 LPG 누출 탐지 모델 개발을 통해 센서 데이터를 기반으로 누출 여부를 효과적으로 예측할 수 있었다.
- 랜덤 포레스트 모델이 가장 높은 성능을 보여 최종 모델로 선정되었다.

b. 개발 의의

- 실시간 탐지를 통해 사고 예방 및 작업자 안전을 보장할 수 있다.
- 비용 효율적인 유지보수 및 관리가 가능해질 것으로 기대된다.
- 데이터 기반 의사 결정이 가능해져 기존의 수동적 탐지 시스템을 대체할 수 있다.

c. 한계 및 제안

- 데이터가 제한적이어서 모델의 일반화 성능을 평가하는 데 한계가 있었다.
- 실제 현장의 복잡한 환경 변수를 반영하기 위해 추가적인 데이터 확보가 필요하다.
- 다양한 환경에서 모델을 테스트하여 추가적인 모델 최적화를 진행해야 한다.