LPG 가스 누출 여부 분석 시스템

2318038 이예준

1. 개발의 목적

1.1. 개발의 의의

LPG는 화학적 성질상 인화성과 폭발성이 높아 누출 시 대규모 화재 및 인명 사고로 이어질 위험이 큽니다. 머신 러닝 모델을 활용해 LPG 가스 누출 여부를 사전에 탐지함으로써 사고를 미연

에 방지할 수 있습니다.

또한 IoT 센서를 활용하여 실시간으로 이루어지는 데이터 수집과 머신 러닝 모델의 분석은, 현장에서 발생하는 가스 누출 문제를 즉각적으로 탐지할 수 있습니다.

머신 러닝은 단순히 LPG 누출 여부 이외에도 평소 존재하는 안전한 수준의 LPG 농도와 위험 수준의 누출을 구분하는 데 도움을 줍니다. 이를 통해 불필요한 경보를 줄이고 운영 비용을 절감할 수 있습니다.

1.2. 머신 러닝 모델 활용 대상

산업현장: 화학 공장, LPG 저장소 및 충전소, 석유 및 가스 처리 플렌트 등

주거 및 상업용 시설: 가정 및 LPG가 사용되는 상업시설 등

운송 및 물류 산업: LPG 운송 차량, 가스 연료를 사용하는 선박 및 항공기 등

2. 개발 내용

2.1. 데이터에 대한 구체적 설명

데이터는 1001개의 표본과 9개의 환경적 속성을 포함하고 있습니다.

독립변수: Alcohol, CH4, CO, H2, LPG, Propane, Smoke, Temp

종속변수: LPG_Leakage (LPG 가스 누출 여부)

2.2. 사용할 성능 지표

모델 성능을 평가하기 위해 정확도, 정밀도, 재현율을 사용할 예정입니다.

모델을 훈련시킨 후, 교차 검증을 사용하여 모델의 성능을 검증할 예정입니다.

데이터를 훈련 세트와 테스트 세트로 분할한 후, 테스트 세트에서 모델을 평가하여 실제 성능을 측정합니다. 교차 검증을 통해 모델의 일반화 성능을 보다 정확하게 평가할 수 있습니다.

3. 개발 과정

3.1. 라이브러리 설치, 데이터 로드, 한글 폰트 설정

프로젝트 진행을 위해 필요한 라이브러리들을 설치합니다.

데이터를 csv 파일 형식으로 저장하고 이를 pandas 를 사용하여 불러옵니다.

그래픽에서 한글이 깨지지 않도록 한글 폰트를 설정합니다.

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# 한글 폰트 설정 (옵션)
plt.rcParams['font.family'] = 'Malgun Gothic'  # Windows: 맑은 고딕
plt.rcParams['axes.unicode_minus'] = False  # 마이너스 기호 깨짐 방지

# 데이터 로드 및 확인
file_path = './data/7.lpg_leakage.xlsx.csv'  # 데이터 파일 경로
data = pd.read_csv(file_path)  # CSV 파일 읽기
```

3.2. 데이터 전처리

데이터의 구조와 결측값을 확인하였고 데이터에 결측값이 있는 경우 결측값이 있는 행을 삭제하였습니다.

Lpg_leakage 를 종속변수로 설정하고 나머지 열들을 독립 변수로 설정한 후 모델 학습에 사용할 데이터를 준비하였습니다.

```python
# 데이터 구조 확인
print(data.head())
print(data.info())

# 데이터 요약: 통계 정보 확인
print(data.describe())

# 결측값 처리: 결측값이 있다면 제거
data = data.dropna()
```

3.3. 데이터 시각화

Lpg_leakage 의 값에 따른 데이터 분포를 시각화 하여 누출 여부가 얼마나 균등하게 분포되어 있는지 확인합니다.

상관행렬을 통해 각 변수들 간의 상관관계를 시각화 하여 특정 변수들의 연관성을 파악합니다.

```python
# 데이터 분포 시각화
plt.figure(figsize=(12, 8))
sns.countplot(x='LPG_Leakage', data=data)
plt.title('LPG 가스 누출 여부 분포')
plt.show()

# 변수 간 상관관계 시각화 (히트맵)
corr_matrix = data.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('변수 간 상관관계')
plt.show()
```

3.4 데이터 분할 및 스케일링

종속변수(LPG_leakage)와 종속변수(LPG_leakage 제외 나머지)를 분리합니다.

Standarscaler 를 사용하여 데이터를 표준화합니다. 표준화는 각 변수의 평균을 0, 표준편차를 1 로 만들어 모델 학습을 효율적으로 만듭니다.

```
# 독립 변수(X)와 종속 변수(y) 분리
X = data.drop(columns=['LPG_Leakage'])  # 독립 변수
y = data['LPG_Leakage']  # 종속 변수

# 데이터 스케일링: 표준화
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 데이터 분할: 훈련 데이터와 테스트 데이터로 분리 (80% 훈련, 20% 테스트)
X_train, X_test, y_train, y_test = train_test_split( *arrays: X_scaled, y, test_size=0.2, random_state=42, stratify=y)
```

3.5 랜덤 포레스트 분류 모델 학습

랜덤 포레스트 모델을 사용하여 훈련 데이터를 학습합니다.

```
# 머신러닝 모델: 랜덤 포레스트 분류기
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

3.6. 모델 평가

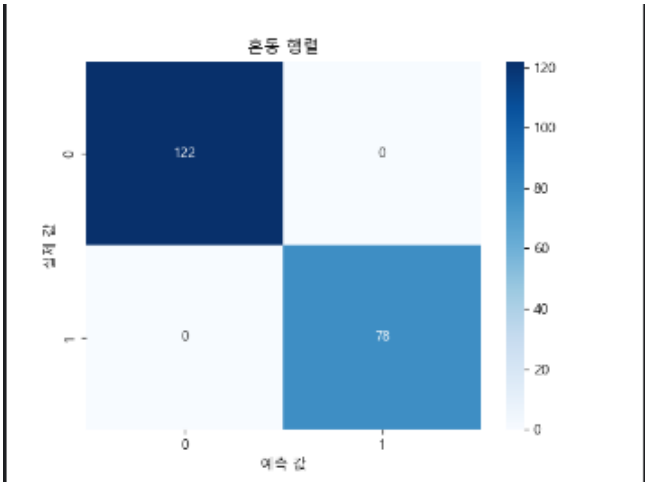
정밀도, 재현율, F1-score 등을 포함한 모델 성능 보고서를 출력합니다.

혼동 행렬 계산 및 시각화를 통하여 모델의 예측 결과와 실제 값의 차이를 시각화 하여 모델의 성능을 직관적으로 이해할 수 있습니다.

```
# 모델 성능 평가
accuracy = accuracy_score(y_test, y_pred)
print(f"모델 정확도: {accuracy:.2f}")

# 분류 보고서 출력
print("\n분류 보고서:")
print(classification_report(y_test, y_pred))

# 혼동 행렬 시각화
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No Leak', 'Leak'], yticklabels=['No Leak', 'Leak'])
plt.title("혼동 행렬")
plt.xlabel("예측 값")
plt.ylabel("실제 값")
plt.show()

# 피처 중요도 시각화 (랜덤 포레스트의 피처 중요도)
feature_importance = model.feature_importances_
features = X.columns
plt.figure(figsize=(10, 6))
sns.barplot(x=feature_importance, y=features)
plt.title('특성 중요도')
plt.show()
```

4. 결과

4.1. 데이터 시각화

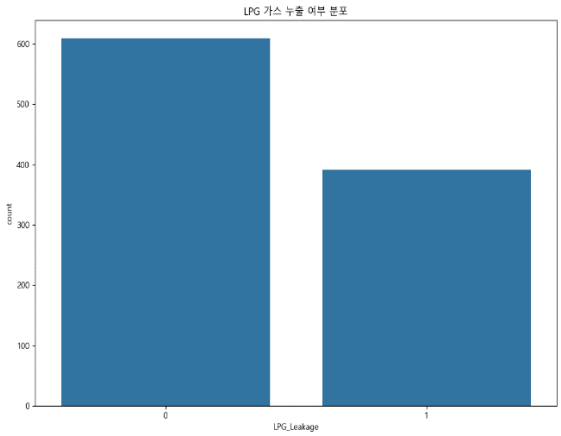


True Negative: 122 건

True Positive: 78 건

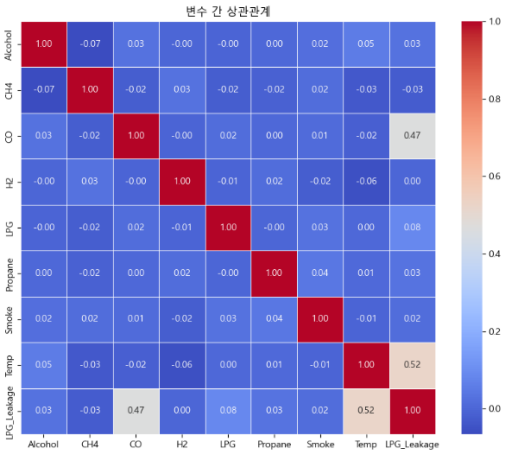
False Positive: 0 건

False Negative: 0 건



LPG 가스 누출 없음: 약 600 건

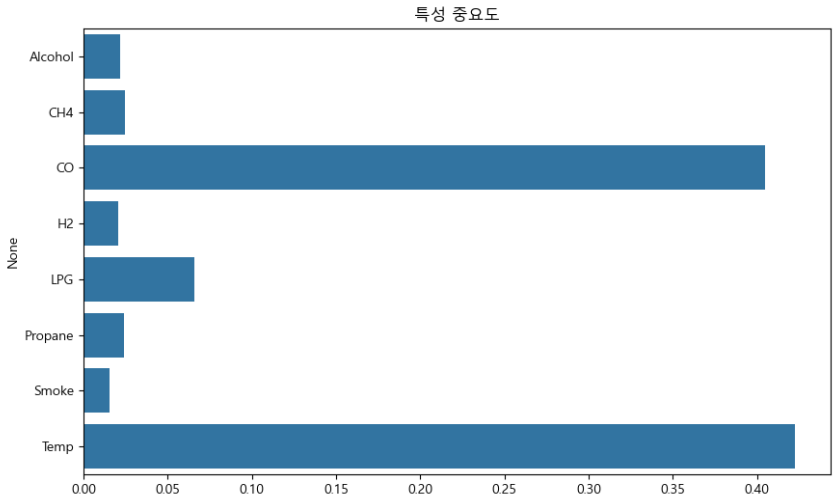
LPG 가스 누출 있음: 약 400 건



CO: 가장 높은 상관관계를 보입니다. CO 농도가 높아질수록 LPG 가스 누출 가능성이 올라갑니다.

TEMP: 온도와의 상관관계도 높은 편에 속합니다.

ALCOHOL. CH4, H2, LPG, Propone, Smoke: 상관관계가 낮거나 거의 없음



CO 와 TEMP 가 높은 중요도를 보입니다.

4.2. 개발 후 느낀 점

모델의 성능을 높이기 위해서는 데이터 전처리 과정이 중요하다는 것을 느꼈습니다. 결측값의 처리, 데이터 스케일링, 적절한 특성 선택이 모델의 정확도를 높일 수 있다는 것을 알게 되었습니다.

다양한 머신러닝 모델 중 랜덤 포레스트를 선택하여 과적합을 방지하고 안정적인 예측 성능을 보였습니다. 이처럼 데이터의 특성과 문제의 성격에 맞는 방식으로 모델을 선택하는것의 중요성을 느꼈습니다.