머신 러닝 모델을 이용한 LPG 가스 누출 탐지기

LGG(LPG Gas Guard)

6 조

제출일자 : 2024.08.27

발표일자 : 2024.08.28

조원 : 김민석(조장),이승연, 임하영, 장영수

목차

1. LPG 가스 누출 탐지 프로그램 개요
2. 개발 목적
3. 배경지식
    a. LPG 가스
    b. 머신 러닝
    c. 회귀분석
    d. 선형 회귀
4. 사례
    a. 평창 LPG 가스 충전소 폭발 사고
5. 개발 내용
    a. HEATMAP 을 통한 데이터간 상관관계 분석
    b. 데이터에 대한 설명
    c. 해당 데이터를 판단하기 위한 머신 러닝 모델 선정
    d. 사용한 성능 지표 종류
6. 개발 결과
    a. 로지스틱 회귀 모델 머신 러닝 결과 및 성능 지표 결과
    b. 의사결정 나무 모델 머신 러닝 결과 및 성능지표 결과
    c. 머신 러닝 모델의 성능 결과
7. 결론

머신 러닝 모델을 이용한 LPG 가스 누출 탐지기(LGG)

발표자 : 이승연

팀원 : 김민석, 이승연, 임하영, 장영수

Github : minseok3/team6_pjt- (github.com)

1. LPG 가스 누출 탐지 프로그램 개요

a. LPG(액화석유가스)는 가정과 산업에서 광범위하게 사용되지만, 가스 누출 시 심각한 화재와 폭발 사고를 초래할 수 있다. 최근 LPG 가스 누출로 인한 사고가 빈번하게 발생하면서, 이를 예방하고 안전성을 높이기 위한 시스템 개발의 필요성이 커지고 있다.

b. 이에 따라, 우리는 Python 을 활용하여 로지스틱 회귀 모델과 의사결정 나무 모델을 기반으로 LPG 가스 누출 여부를 정확하게 판단하는 프로그램을 개발하였다. 해당 프로그램으로 가스 누출을 조기에 탐지하고 신속한 대응을 통해 안전성을 높일 수 있다.

2. 개발 목적

a. 본 프로그램의 주요 목적은 LPG 가스 누출을 신속하고 정확하게 탐지하여 안전사고를 예방하는 것에 있다. LPG 가스는 무색, 무취의 특성으로 인해 누출 시 감지가 어렵고, 이로 인해 심각한 화재와 폭발 사고가 발생할 수 있다.

b. 로지스틱 회귀 모델과 의사결정 나무 모델을 활용하여 가스 누출의 가능성을 신속하게 분석하고 조기에 탐지한다.

c. 다양한 데이터 입력 값을 바탕으로 LPG 가스 누출 여부를 정확하게 판단하여 오탐지와 탐지 누락을 최소화한다.

d. 가정과 산업 현장에서의 안전성을 높이고, 가스 누출로 인한 화재 및 폭발 사고를 예방하여 인명과 재산을 보호한다.

e. 실시간으로 데이터를 분석하고 결과를 제공하여, 신속한 대응과 문제 해결을 지원한다.

3. 배경지식

a. LPG 가스 : 연료와 화학 연료로 널리 사용되며, 산업현장, 상업시설, 가정 등 다양한 분야에서 사용한다. 무색, 무취로 일반적으로 감지하기 어렵기 때문에 누출이 발생할 경우 조기에 발견하지 못하면 대규모 폭발 및 화재를 야기한다.
b. 머신 러닝(Machine Learning) : 종속 변수와 하나 이상의 독립 변수 간의 관계를 설명하고, 예측하는데 사용한다.
c. 회귀분석(Regression Analysis) : 데이터 간의 관계를 모델링하고 학습함으로써 미래의 값을 예측하고, 특정 변수가 다른 변수에 미치는 영향을 파악하는 통계적 기법이다.
d. 선형 회귀(Linear Regression) : 독립 변수와 종속 변수 간의 관계를 선형으로 모델링한다.

4. 사례

a. 평창 LPG 가스 충전소 폭발 사고
   i. 사고 일시 : 2024 년 1 월 1 일 오후 8 시 40 분경
   ii. 사고 장소 : 강원도 평창군 용평면 장평리
   iii. 사고 발생 전 상황 : 충전소에서 가스가 새어 바닥에 깔리고 있다는 신고가 119 에 접수
   iv. 사고 발생 : 신고 접수 후 약 20 분 후 폭발 사고 발생
   v. 사고 원인 : 탱크로리를 통해 가스를 이충전하는 과정에서 LPG 가스가 누출되어 주변으로 확산되었고, 원인 미상의 점화 원인에 의해 폭발.
   vi. 피해 상황 : 주민 5 명 부상(1 명 사망), 주택 건물 14 채 파손, 차량 14 대 피해
   vii. 후속 조치 : 사고 발생 직후 현장 안전 조치 및 피해 주민 구조, 원인 조사 및 재발방지 대책 마련

5. 개발 내용

a. HEATMAP 을 통한 데이터간 상관관계 분석
   i. HEATMAP 을 사용해서 LPG 가스 누출 탐지에 가장 큰 영향을 미치는 변수 값이 Temp, CO 라는 것을 파악했다. (그림 1. 참조)
   ii. 우측을 막대바를 통해 누출(1)에 가까울수록 빨간색을, 누출 없음(0)에 가까울수록 파란색을 표시하고 있음을 확인할 수 있다. HEATMAP 을 분석해

보면, CO 와 Temp 의 값이 각각 0.47 과 0.52 로 나타나며 이 두가지 변수가 빨간색(1)에 가까운 값을 보인다. 이에 따라 LPG 가스 누출 여부 판단에 있어 CO 와 Temp 가 중요한 영향을 미친다는 것을 확인할 수 있다.

iii. 하지만 유해가스 측정기가 여러 가지의 가스 값을 측정하는 점을 고려하고, 현장 상황에 맞추어 두가지 변수 값만으로는 LPG 가스 누출 여부를 확인하기에는 한계가 있다고 판단하였다. 따라서 6 개의 추가 가스 값을 포함하여 총 8 개의 가스와 요소를 활용한 데이터 분석을 진행하였다.

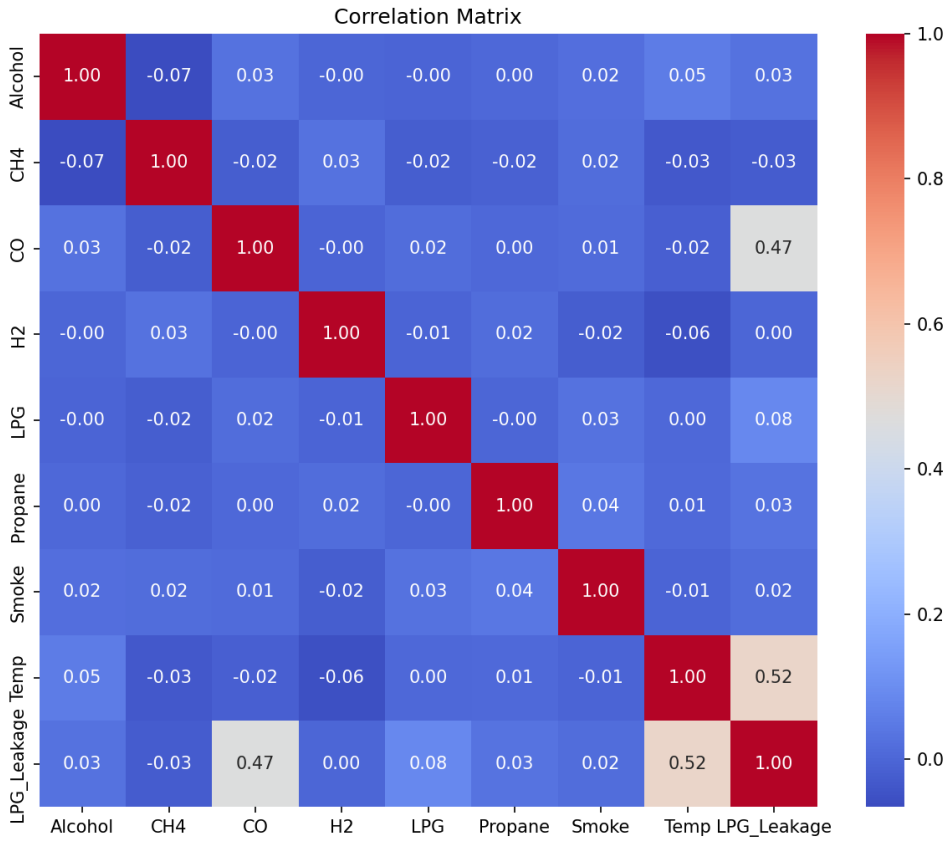

그림 1 HEATMAP 시각화 자료

b. 데이터에 대한 설명

i. 데이터의 독립변수(속성) : 8 가지의 각종 가스 및 요인 (Alcohol, CH4, CO, H2, LPG, Propane, Smoke, Temp)

1. Alcohol : 높은 농도로 존재하면, 다른 가스들과 혼합될 때 그 상호작용이 LPG 측정에 영향을 줄 수 있다.
2. CH4 : 메탄은 LPG 는 모두 탄화수소 계열로, 메탄 농도가 높으면 LPG 및 가스 혼합물의 폭발에 영향을 줄 수 있다.

3. CO : 일산화탄소의 농도가 높다면 LPG 및 다른 가스의 연소를 야기한다.
4. H2 : 수소와 LPG 가 동시에 존재하는 경우, 두 가스 모두 폭발 위험을 야기한다.
5. LPG : 프로판과 부탄의 혼합물로 상온에서 액체 상태로 저장되지만, 누출 시 기체 상태로 변해 공기 중으로 퍼진다.
6. Propane : 프로판은 LPG 의 주요 성분으로 두 값이 비례관계를 가져 프로판의 농도가 높으면 LPG 의 농도도 높을 가능성이 높다.
7. Smoke : 불완전한 연소를 나타나며, 연기가 감지되면 LPG 또는 다른 연료가 제대로 연소되지 않고 있음을 나타냄. 즉, 연기의 존재는 타 가스의 연소를 야기한다.
8. Temp : 온도가 높으면 LPG 가 기화하여 공기 중에 퍼질 가능성이 증가한다.

ii. 데이터의 종속변수 : LPG 가스 누출 여부(LPG-Leakage)

iii. 총 데이터 개수 1000 개 (학습 데이터 800 개, 테스트 데이터 200 개)

iv. 데이터 수집 및 전처리

1. 데이터 수집 : 데이터는 CSV 파일 형식으로 제공되었고, 총 8 개의 독립변수에 관한 데이터 1000 개를 바탕으로 가스 및 온도 측정값으로 LPG 가스의 누출 여부 확인한다.
2. 데이터 전처리 : 데이터 분석과 모델 학습의 품질을 높이기 위해 데이터 정리와 변환을 수행하는 과정이다. 다음과 같은 순서로 실행했다.
    a. 데이터 준비 - array 를 사용해 8 개의 독립변수를 가진 데이터를 배열(list)형태로 표현한다.
    b. 데이터 정규화 – MinMaxScaler 를 사용해 1000 개의 데이터를 0 과 1 사이의 값으로 조정해 모든 데이터를 일정한 범위로 변환한다.

c. 해당 데이터를 판단하기 위한 머신 러닝 모델 선정

i. 로지스틱 회귀(Logistic Regression) 모델 선정 :

해당 데이터의 종속 변수(0 or 1)와 같은 이진분류 문제를 해결하는 데 사용하는 통계적 모델로 LPG 가스의 누출 또는 정상을 판단하기에 적합하다고 생각해 사용하였다.

ii. 의사결정 나무(DecisionTreeClassfier) 모델 선정 :

해당 데이터의 독립변수 값처럼 비 선형적이고 복잡한 데이터 패턴을 효과적으로 모델링할 수 있다. 다양한 특성 간의 비선형 관계를 포착해 데이터의 복잡성을 반영하는 것이 단순 데이터를 판단하는데 유리한 로지스틱 회귀 모델과 비교를 위해 사용하였다.

d. 사용한 성능 지표 종류

i. 정확도(Accuracy)

1. 전체 데이터 중에서 올바르게 분류된 샘플의 비교를 나타내는 성능 지표로 모델의 전반적인 성능을 간단명료하게 이해할 수 있어서 사용하였다.

ii. ROC-Curve

1. 원래 신호 감지 이론에서 사용되던 개념이지만 현재는 머신 러닝의 성능을 평가하는데 사용된다. ROC-Curve 는 다양한 임계 값에 대한 성능을 시각화 하여 곡선이 왼쪽 상단에 가까울수록 모델의 성능이 우수하다는 것을 한눈에 판단할 수 있어 사용하였다.

iii. 성능 지표 선정 이유

1. 머신 러닝 모델이 LPG 가스 누출 상황을 정확하게 판단하는 것이 가장 중요하다. 만약 모델이 가스 누출이 발생했음에도 불구하고 이를 감지하지 못하거나 '없다'고 판단하면, 심각한 안전 위험을 초래할 수 있다.
2. 모델의 정확도가 높아야만 이러한 위험을 최소화하고 사고를 사전에 예방할 수 있다고 생각하여 'Accuracy', 'ROC-Curve' 성능 지표를 선정해 모델이 얼마나 신뢰성 있게 가스 누출을 감지하는지 평가하고자 하였다.

6. 개발 결과

a. 로지스틱 회귀 모델 머신 러닝 결과 및 성능 지표 결과

i. 머신 러닝 결과 및 성능 지표 결과

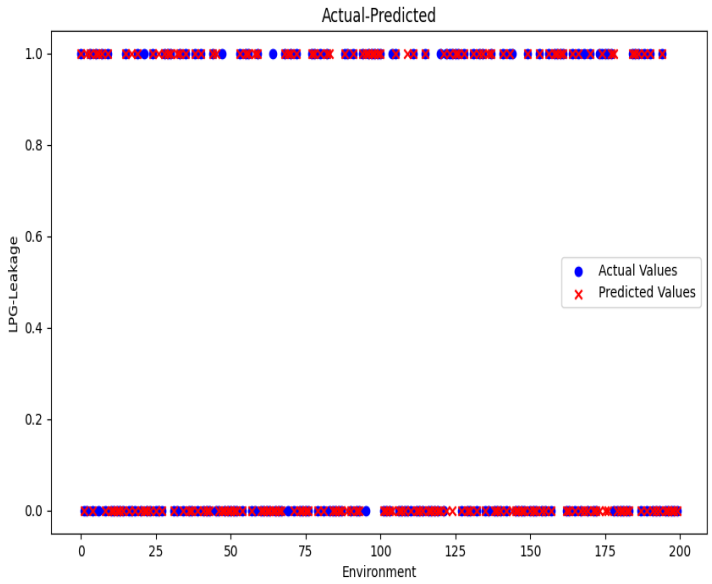

그림 2 로지스틱 회귀 머신 러닝 결과

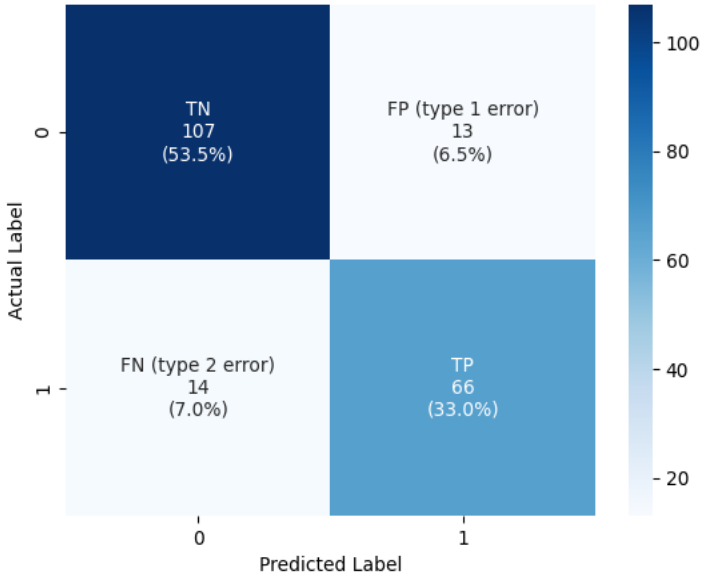

그림 3 로지스틱 회귀 머신 러닝 성능지표 Accuracy 결과값

1. 그림 2 는 로지스틱 회귀 모델 머신 러닝 결과이다. 이 그래프에서 X 축은 8 개의 독립 변수로 구성된 1000 개의 데이터 중 20%를 분할하여 얻은 200 개의 테스트 데이터를 나타낸다. Y 축은 LPG 누출 여부를 1(누출)과 0(누출 없음)으로 나타낸다. 로지스틱 회귀 모델을 적용한 결과, 실제 값(파란색 원형 점)과 예측 값 (빨간색 X 형 점)이 86.5%의 정확도로 일치함을 그림 3 을 통해 확인할 수 있다.
2. 그림 3 은 머신 러닝 성능지표 Accuracy 로 로지스틱 회귀 모델 머신 러닝 결과값을 시각화 한 것이다. 이 그래프에서 X 축은 예측 값(1: 누출, 0:누출 없음)을 Y 축은 실제 값 (1:누출, 0:누출 없음)을 나타낸다.
   a. True Positive(TP) : 실제 누출이 발생한 경우를 정확히 누출로 예측한 비율
   b. True Negative(TN) : 실제 누출이 발생하지 않은 경우를 정확히 누출 없음으로 예측한 비율
   c. False Positive(FP) : 실제 누출이 없지만 누출이 있다고 잘못 예측한 비율
   d. False Negative(FN) : 실제 누출이 발생했지만 누출이 없다고 잘못 예측한 비율

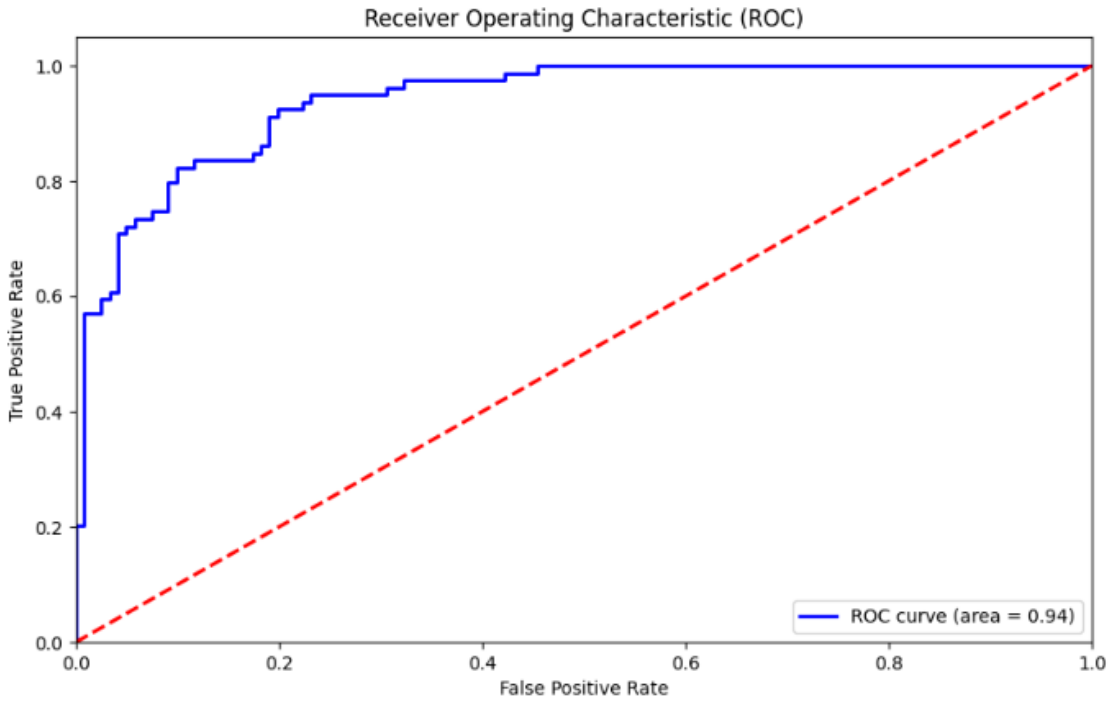

그림 4 로지스틱 회귀 머신 러닝 성능 지표 ROC-Curve 결과값

3. 그림 4 는 로지스틱 회귀 모델 머신 러닝 성능 지표 ROC-Curve 를 시각화 한 것이다. 이 그래프에서 X 축은 False Positive Rate(FRR)을 나타내며 이는 실제로 부정적인 사례 (누출이 없는 경우)인데 모델이 누출로 잘못 예측한 경우를 의미한다. Y 축은 True Positive Rate(TPR)을 나타내며 이는 실제로 긍정적인 사례 (누출이 있는 경우)로 모델이 누출로 정확히 예측한 경우를 의미한다.
4. 그림 4 의 시각화 자료를 보면 ROC 곡선이 계단 모양을 보이고 있으며, 이는 모델의 예측 성능이 일정 구간에서는 변동이 크지 않음을 나타낸다. 따라서 모델의 정확도가 비교적 높지 않다고 해석할 수 있다.

b. 의사결정 나무 모델 머신 러닝 결과 및 성능지표 결과

i. 머신 러닝 결과 및 성능 지표 결과

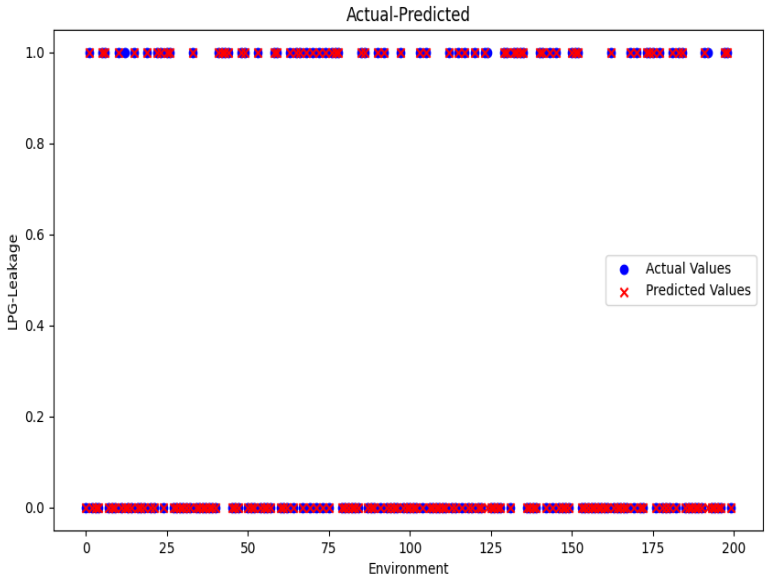

그림 5 의사결정 나무 모델 머신 러닝 결과

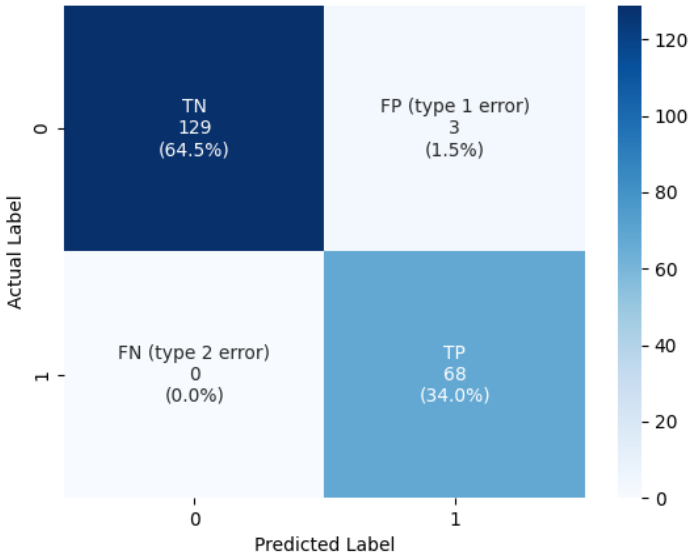

그림 6 의사결정 나무 모델 머신 러닝 성능지표 Accuracy 결과값

1. 그림 4 는 의사결정 나무 모델 머신 러닝 결과이다. 이 그래프에서 X 축은 8 개의 독립 변수로 구성된 1000 개의 데이터 중 20%를 분할하여 얻은 200 개의 테스트 데이터를 나타낸다. Y 축은 LPG 누출 여부를 1(누출)과 0(누출 없음)으로 나타낸다. 로지스틱 회귀 모델을 적용한 결과, 실제 값(파란색 원형 점)과 예측 값 (빨간색 X 형 점)이 98.5%의 정확도로 일치함을 그림 5 를 통해 확인할 수 있다.
2. 그림 5 은 머신 러닝 성능지표 Accuracy 로 의사결정 나무 모델 머신 러닝 결과값을 시각화 한 것이다. 이 그래프에서 X 축은 예측 값(1:누출, 0:누출 없음)을 Y 축은 실제 값 (1:누출, 0:누출 없음)을 나타낸다.
   a. True Positive(TP) : 실제 누출이 발생한 경우를 정확히 누출로 예측한 비율
   b. True Negative(TN) : 실제 누출이 발생하지 않은 경우를 정확히 누출 없음으로 예측한 비율
   c. False Positive(FP) : 실제 누출이 없지만 누출이 있다고 잘못 예측한 비율
   d. False Negative(FN) : 실제 누출이 발생했지만 누출이 없다고 잘못 예측한 비율

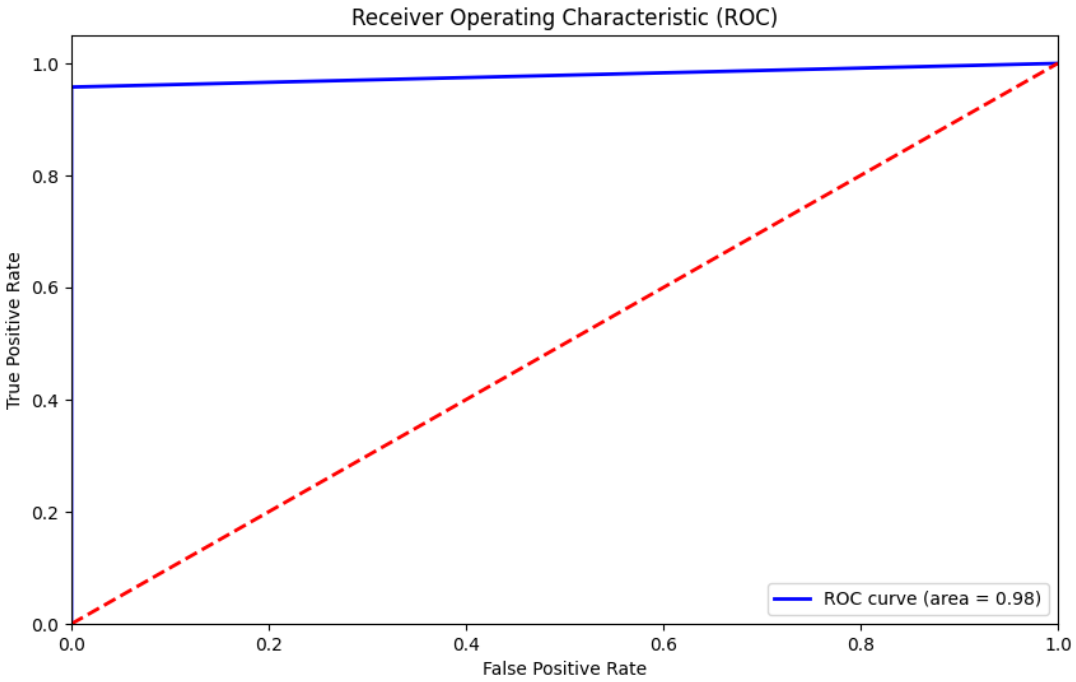

그림 7 의사결정 나무 모델 머신 러닝 성능 지표 ROC-Curve 결과값

3. 그림 4 는 의사결정 나무 모델 머신 러닝 성능 지표인 ROC-Curve 를 시각화 한 것이다. 이 그래프에서 X 축은 False Positive Rate(FRR)을 나타내며 이는 실제로 부정적인 사례 (누출이 없는 경우)인데 모델이 누출로 잘못 예측한 경우를 의미한다. Y 축은 True Positive Rate(TPR)을 나타내며 이는 실제로 긍정적인 사례 (누출이 있는 경우)로 모델이 누출로 정확히 예측한 경우를 의미한다.
4. 그림 7 의 시각화 자료를 보면 ROC 곡선이 좌상단에 가까워지고 있는 것을 볼 수 있다. 이는 모델의 예측 성능이 상대적으로 좋다는 것을 나타낸다. 따라서 모델의 정확도가 비교적 높다고 해석할 수 있다.

ii. 머신 러닝 모델 간 성능 비교

1. 로지스틱 회귀 모델, 의사결정 나무 모델을 사용했을 시 각각의 머신 러닝 결과를 다양한 시각화 자료를 통해 확인하였다. 로지스틱 회귀 모델의 정확도는 86.5%, 의사결정 나무 모델의 정확도는 98.5%로 의사결정 나무 모델이 성능이 좋다는 것을 확인할 수 있었다.

c. 머신 러닝 모델의 성능 결과

i. 8 개의 독립변수 당 800 개의 학습 데이터와 200 개의 테스터 데이터를 사용하여 두 모델의 머신 러닝 결과만을 봤을 때는 차이를 확인하기 어려웠으나, 성능 지표를 교차 검증하여 두 머신 러닝 간 정확도가 약 13% 차이를 보여 유의미한 결과를 확인할 수 있었다. 따라서 의사결정 나무 모델이 로지스틱 회귀 모델에 비해 월등한 정확도를 가지고 있다는 결과를 도출하였다.

7. 결론

a. 결론

i. LPG 가스 누출 탐지를 위한 의사결정 나무 머신 러닝 모델과 그 결과에 대한 성능 지표 분석은 100%에 가까운 정확도를 도출해냈다. 현재는 1000 개의 가상 데이터를 분석하였으나 실제 현장에서는 유해가스탐지기 등을 이용해 실시간으로 정보를 입력 받고 그 정보를 바탕으로 LPG 가스 누출을 탐지할 수 있는 프로그램으로 응용 가능할 것이다.