Flame Watch AI

이름: 김도영

학번: 2118004

Github: https://github.com/red21-22/01_Python_and_Git/blob/main/Final%20Test.py

1. 안전 관련 머신러닝 모델 개발 관련 요약

산불은 생태계와 인류에 막대한 피해를 주는 자연재해로, 사전 감지와 예방이 중요하다. 본 프로젝트는 기상 데이터를 기반으로 머신러닝 모델을 활용하여 산불 발생 가능성을 예측하고자 한다. 랜덤 포레스트를 주요 모델로 사용하여 산불 감시의 정확도를 높이고, 효과적인 대응 방안을 마련하는 데 기여하고자 한다.

2. 개발 목적

a.머신러닝 모델 활용 대상

- 불은 자연환경과 인명, 재산에 막대한 피해를 끼치는 재난 중 하나로, 조기 감지가 피해 최소화의 핵심 요소로 강조되고 있다. 전통적인 감시 방식은 인적 자원에 크게 의존하며, 넓은 지역을 효과적으로 감시하기에는 한계가 있다.

- 이러한 문제를 해결하기 위해 머신러닝 기술이 각광받고 있으며, 빅데이터와 인공지능을 활용한 산불 감시 시스템은 기존 방식의 효율성을 크게 향상시킬 잠재력을 가지고 있다. 본 보고서는 머신러닝 기반 산불 감시 시스템 개발을 통해 산불의 조기 감지와 예측 가능성을 높이고, 효과적인 대응 체계를 마련하고자 해당 프로젝트를 준비했다.

b . 개발의 의의

- 산불 감시의 자동화 및 정밀화를 통해 인명과 환경 보호에 기여하기 위한 것으로, 산불 조기 감지로 피해를 예방하는 것은 물론, 산림 관리 효율성을 높여 공공 안전을 강화 및 경제적 손실을 최소화하는 데 중요한 가치를 창출할 것이다. 또한, 빅데이터를 활용한 환경 분석 기술 발전에 기여하여 관련 연구 및 응용 분야의 확장을 도모할 수 있다.

c. 데이터의 어떠한 독립 변수를 사용하여 어떠한 종속 변수를 예측하는지

독립 변수 : 기상 데이터(온도, 습도, 풍속 등), 토양 수분, 위성 및 항공 이미지의 화염 및 연기 특징, 지형 정 보(고도, 경사도 등), 산림 밀도, 계절적 요인 등

종속 변수 : 산불 발생 가능성(이진 분류: 발생/비발생), 산불 발생 지역의 위험 등급(다중 클래스 분류) 산불 확산 예상 범위(회귀 분석)

3. 배경지식

a. 데이터 관련 사회 문제 설명

산불은 전 세계적으로 발생 빈도와 강도가 증가하며 심각한 사회적 문제로 대두되고 있다. 특히 기후 변화로 인해 건조한 환경과 극단적인 기상 조건이 늘어나면서 산불 발생 가능성이 높아지고 있는 상황이다. 특히 산불은 자연 생태계를 파괴하고, 대기 오염을 유발하며, 경제적 손실과 인명 피해를 초래하는데, 이러한 문제를 해결하기 위해 조기 감지 및 효과적인 대응이 필수적이나, 기존의 감시 시스템은 한정된 자원으로 인해 특히 우리나라 같은 수 많은 산림 지역을 실시간으로 관찰하는 데 한계가 있다.

그러나 데이터 분석 기술의 발전은 이러한 문제를 해결할 수 있는 가능성을 열어준다. 기상 데이터, 위성 및 드론 촬영 데이터, IoT 센서 데이터를 활용하면 산불 위험 요소를 실시간으로 모니터링하고 예측할 수 있지만,
이러한 데이터는 대용량이고 복잡한 특성을 가지므로, 이를 효과적으로 처리하고 유의미한 정보를 추출하기 위해 고도화된 분석 방법이 필요하다.

b. 머신러닝 모델 관련 설명

머신러닝은 데이터를 기반으로 패턴을 학습하여 예측 및 분류 작업을 자동화하는 기술인데, 산불 감시 시스템에서는 다음과 같은 머신러닝 모델이 활용될 수 있다.

1) 지도 학습(Supervised Learning): 산불 발생 여부를 예측하거나 위험 등급을 분류하는 데 사용된다. 예를 들어, 과거 산불 발생 데이터와 기상 조건을 학습하여 향후 산불 발생 가능성을 예측한다.

2) 비지도 학습(Unsupervised Learning): 산림 데이터의 클러스터링을 통해 위험 지역을 파악하거나 산림 유형을 분류하는 데 활용된다.

3) 딥러닝(Deep Learning): 위성 이미지나 드론 영상에서 연기, 불꽃, 열원 등의 특징을 자동으로 인식하여 실시간 경보를 제공하는 데 적용된다.

4) 강화 학습(Reinforcement Learning): 자원을 효율적으로 배치하거나 대응 전략을 최적화하는 데 사용될 수 있다.

이러한 머신러닝 기술은 데이터의 복잡성을 처리하고, 기존 시스템보다 높은 정확성과 실시간 대응 능력을 제공한다. 특히 빅데이터와 결합하여 다차원적 분석을 수행함으로써 산불 감시 및 관리의 효율성을 극대화할 수 있다.

4. 개발 내용

a. 데이터에 대한 구체적 설명 및 시각화

i) 데이터 개수, 데이터 속성 등

사용된 데이터셋은 총 1,000 개의 샘플로 구성되어 있으며, 각 샘플은 4 개의 독립 변수와 1 개의 종속 변수를 포함하고 있는데, 주요 속성은 다음과 같다.

- 온도 (temperature): 15 도에서 45 도 사이의 값으로, 산불 발생에 영향을 미칠 수 있는 환경적인 변수다.
- 습도 (humidity): 10%에서 90% 사이의 값으로, 건조한 환경이 산불 발생에 영향을 줄 수 있다.
- 풍속 (wind_speed): 0m/s 에서 20m/s 사이로, 강한 바람은 산불을 확산시킬 수 있다.
- 토양 습도 (soil_moisture): 5%에서 50% 사이의 값으로, 토양의 습도는 산불의 발생 가능성과 연관이 있을 수 있다.
- 산불 발생 여부 (fire_risk): 목표 변수로, 값은 0(비발생) 또는 1(발생) 이다.

ii) 데이터 간 상관관계 설명 등

- 온도와 습도는 일반적으로 음의 상관관계를 가질 수 있는데, 온도가 높을수록 습도는 낮아지는 경향이 있다.
- 풍속과 산불 발생은 양의 상관관계를 보일 가능성이 크다. 특히 바람이 강할수록 산불이 확산될 가능성이 높다.
- 토양 습도는 산불 발생에 대한 완충 역할을 할 수 있는데, 토양 습도가 높을수록 산불 발생 확률이 낮아질 수 있다.

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

i) 독립변수, 종속변수 설정

- 독립변수 (X): 모델의 입력으로 사용되는 변수는 온도, 습도, 풍속, 토양 습도다. 이 변수들은 산불 발생에 영향을 미칠 수 있는 환경적인 요소들이다.
- 종속변수 (y): 예측하려는 변수는 산불 발생 여부 (fire_risk)다.. 이 변수는 0(비발생) 또는 1(발생)로 정의되어 있으며, 모델은 이 변수의 값을 예측한다.

c. 머신러닝 모델 선정 이유

i) 설명한 데이터를 기반으로 머신러닝 모델 선정 이유 설명

주어진 데이터는 환경 변수들을 바탕으로 산불 발생 여부를 예측하는 분류 문제인데, 랜덤 포레스트 분류기를 선택한 이유는 다음과 같다.

- 랜덤 포레스트는 다수의 결정 트리를 결합하여 예측 성능을 향상시킬 수 있는 앙상블 학습 방법이다.
- 과적합을 방지하고 변수 중요도를 분석할 수 있어, 데이터에서 중요한 특성을 쉽게 파악할 수 있다.
- 산불 발생 여부는 이진 분류 문제이므로, 랜덤 포레스트는 적합한 모델이다.

ii) 성능 비교를 위한 머신러닝 모델 선정 이유

- 다른 머신러닝 모델들도 함께 평가하여 성능을 비교하고자 하였으며, 예를 들어 로지스틱 회귀, 결정 트리등과 함께 성능을 비교할 수 있는데. 이렇게 다양한 모델을 사용하여 모델 성능을 최적화할 수 있다.

d. 사용할 성능 지표

i) 머신러닝 모델의 성능을 평가하기 위해 사용하는 성능 지표에 관한 설명 등

- 정확도(Accuracy): 전체 샘플 중에서 모델이 정확하게 예측한 비율이다. 하지만 클래스 불균형 문제가 있을 경우 정확도만으로는 충분히 평가할 수 없다.
- 정밀도(Precision): 모델이 산불 발생(1)으로 예측한 샘플 중 실제로 산불이 발생한 비율이다. 산불을 예측하는 데 있어 잘못된 예측을 줄이기 위해 중요하다.
- 재현율(Recall): 실제로 산불이 발생한 샘플 중 모델이 산불 발생으로 예측한 비율이다. 실제 발생한 산불을 놓치지 않기 위해 중요하다.
- F1-score: 정밀도와 재현율의 조화 평균으로, 두 지표의 균형을 맞추고자 할 때 유용하다.

ii) 성능 지표 선정 이유 등

- 산불 발생 예측 데이터 같은 경우는 불균형 데이터 이기 때문에 정확도만으로 모델을 평가하기에는 좀 애매하다. 그러기에 정밀도와 재현율을 종합적으로 고려할 수 있는 F1-score 가 중요한 성능 지표로 선정되었다.

5. 개발 결과

a. 성능 지표에 따른 머신러닝 모델 성능 평가

i) 수치 자료 및 시각화 자료를 사용

랜덤 포레스트(Random Forest) 모델의 성능은 다음과 같은 평가 지표를 통해 분석하였다.

- ◉ 정확도(Accuracy): 85.5%
- ◉ 평균 절대 오차(MAE): 0.12
- ◉ 평균 제곱 오차(MSE): 0.18
- ◉ 루트 평균 제곱 오차(RMSE): 0.42

ii) 다른 머신러닝 모델과 성능 비교

- 랜덤 포레스트(Random Forest) 모델과 로지스틱 회귀(Logistic Regression) 모델을 대상으로 KFold 교차 검증과 다양한 성능 지표를 통해 비교 평가를 진행하였다.

모델	평균 KFold 정확도	MSE	RMSE	MAE	Accuracy
랜덤 포레스트	79.4%	0.211	0.459	0.164	81.0%
로지스틱 회귀	74.2%	0.258	0.508	0.196	76.5%

b. 머신러닝 모델의 성능 결과에 대한 해석

① 랜덤 포레스트

KFold 평균 정확도와 테스트 데이터 정확도가 모두 로지스틱 회귀보다 높게 나타났다. 특히 MSE(평균 제곱 오차), RMSE(제곱근 평균 제곱 오차), MAE(평균 절대 오차) 모두 더 낮은 값을 기록하며, 예측 성능이 우수함을 보여주었으며, 오차 행렬 분석에서도 산불 발생과 비발생 모두 더 높은 정확도를 보여주었다.

② 로지스틱 회귀

상대적으로 낮은 평균 정확도와 높은 MSE, RMSE 를 기록하여 데이터의 비선형적인 특성을 충분히 반영하지 못한 한계가 드러났습니다.

③ 총 해석

랜덤 포레스트는 KFold 교차 검증과 다양한 성능 지표에서 우수한 결과를 보여주어, 산불 감시 시스템에 최적의 모델로 선정되었다. 로지스틱 회귀는 단순한 데이터 구조에서는 효과적일 수 있으나, 본 프로젝트와 같은 복잡한 특성을 포함한 문제에서는 성능이 제한적임을 확인하였다.

3 . 결론

a. 머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명

- 본 프로젝트에서는 랜덤 포레스트 알고리즘을 활용하여 산불 감시를 위한 머신러닝 모델을 개발하였다. 주요 기상 데이터(온도, 습도, 풍속, 토양 습도)를 입력 변수로 사용하여 산불 발생 여부를 예측하였으며, 모델은 79.4%의 정확도와 낮은 오차(MAE: 0.164)를 기록하며 신뢰할 수 있는 성능을 보여주었다.

b. 개발 의의 등

이 모델은 산불 감시에 있어 기상 데이터를 활용한 데이터 기반 의사결정을 가능하게 하는데, 이를 통해 산불 발생 가능성을 사전에 예측함으로써, 인명과 재산 피해를 줄이는 데 기여할 수 있다. 특히, 랜덤 포레스트 모델의 높은 정확도와 효율성은 실시간 산불 감시 시스템에 응용 가능성을 보여준다.

c. 머신러닝 모델의 한계

① 데이터 제한: 모델은 주어진 샘플 데이터로 학습되었으므로, 실제 상황에서의 기상 데이터와 산불 발생 데이터 간의 차이로 인해 성능이 달라질 수 있다.

② 비선형 변수 처리: 랜덤 포레스트는 비선형 변수에 강점을 가지지만, 과도한 상관관계를 가진 변수의 경우 성능 저하를 유발할 가능성이 있다.

③ 실시간 예측 제한: 현재 모델은 학습된 데이터 기반으로 작동하며, 실시간으로 지속적인 데이터를 학습하기 위해 추가적인 알고리즘 구현이 필요하다.