기상 상태에 따른 교통사고 발생 예측

이름: 최민석

학번: 2118064

Github: https://github.com/cms04220/15

1. 안전 관련 머신러닝 모델 개발 관련 요약

날씨 조건에 따른 교통사고 발생여부를 예측하는 머신러닝 모델을 개발하였다. 서울시 2023 년 교통사고 데이터를 기반으로 날씨와 사고 발생 간의 관계를 분석하여, 위험도가 높은 조건에서 교통하고 예방 대책을 마련할 수 있도록 지원하는 모델이다.

2. 개발 목적

a. 머신러닝 모델 활용 대상:

1) 교통 안전 정책 수립을 담당하는 공공기관.

2) 스마트 교통 시스템과 연계된 교통사고 예방 시스템 개발자.

b. 개발의 의의:

1) 날씨 데이터를 통해 교통사고 위험을 사전에 예측하여 사고 예방에 기여.

2) 교통사고로 인한 사회적 비용 감소 및 시민안전증대

3) 정책 수립 시 데이터를 기반으로 한 의사결정 지원

c. 데이터 사용 및 목표

독립변수: 날씨 조건 (맑음, 흐림, 비, 안개, 눈, 기타/불명)

종속변수: 사고 발생 여부(발생(1) 또는 미발생(0))

3. 배경지식

a. 데이터 관련 사회 문제:

1) 악천후에서 교통하고 발생률이 증가하는 것은 잘 알려져 있지만, 구체적으로 어떤 날씨조건이 사고에 큰 영향을 미치는지는 명확히 분석되지 않았다.

2) 데이터 기반으로 사고 위험도를 파악하면, 기상 악화 시의 안전 대책 수립에 기여할 수 있다.

b. 머신러닝 모델 관련 설명:

1) 랜덤 포레스트: 비선형적인 특성을 학습할 수 있는 앙상블 기법으로, 독립변수(날씨)와 종속변수(사고 여부)간의 관계를 효과적으로 모델링할 수 있다. 랜덤 포레스트는 각 독립변수의 중요도를 측정할 수 있어, 날씨 조건별 영향력을 시각적으로 파악할 수 있다.

4. 개발 내용

a. 데이터에 대한 구체적 설명 및 시각화

i. 데이터 개수 및 속성

1. 데이터 출처: 서울시 2023 년 교통사고 통계

2. 데이터개수: 365(1 년간의 일별 데이터)

3. 주요 속성

독립변수: 맑음, 흐림, 비, 안개, 눈, 기타/불명

종속변수: 사고발생여부

ii. 데이터 전처리

1. 결측치가 포함된 행은 삭제 또는 평균값으로 대체

2. 사고발생여부 컬럼은 사고 발생 건수(소계)가 0 보다 큰 경우 1, 아닌 경우 0 으로 이진변환

iii. 데이터 시각화

1. 날씨별 사고 발생 건수

1) 기상 조건별 사고 발생 분포를 막대 그래프로 표현

2) 비, 눈 조건에서 사고 발생률이 상대적으로 높음 확인

2. 상관관계 분석

1) 날씨 조건과 사고 발생 여부 간의 상관관계를 시각화

b. 예측목표

i. 독립변수와 종속변수 설정

독립변수(X): 날씨 데이터(맑음, 흐림, 비, 안개, 눈, 기타/불명)

종속변수(y): 사고 발생 여부(0 또는 1)

c. 머신러닝 모델 선정

랜덤 포레스트 선정 이유:

1) 날씨와 사고 발생 간의 비선형적 관계를 효과적으로 모델링 가능

2) 앙상블 학습으로 과적합 방지와 높은 예측 성능 제공

3) 각 변수의 중요도를 분석하여 날씨 조건별 사고 기여도를 정량적으로 평가 가능

비교 모델:
로지스틱 회귀: 기본적인 선형 모델로 랜덤 포레스트와 성능 비교
결정 트리: 단일 트리 기반으로 간단한 비선형 모델링을 수행

d. 모델 평가

성능 지표:
정확도: 모델의 전체 예측 성능
정밀도: 모델이 사고 발생을 올바르게 예측한 비율
재현율: 실제 사고 발생을 얼마나 잘 예측했는지 평가
F1 스코어: 정밀도와 재현율의 조화 평균
혼동행렬: 예측값과 실제값의 분포를 비교하여 성능분석

모델 성능 결과:
모델 정확도: 약 0.85(85%)
비와 눈 조건에서 사고 발생 예측 정확도가 높았음

e. 특성 중요도 분석

랜덤 포레스트를 통해 각 날씨 조건의 중요도를 분석한 결과:
비와 눈이 사고 발생에 가장 큰 영향을 미침
맑음과 기타/불명 조건은 상대적으로 중요도가 낮음

특성 중요도 시각화:
특성 중요도를 내림차순으로 정렬하여 막대 그래프 생성
중요한 날씨 조건부터 차례대로 정렬하여 분석 결과를 전달

5. 결론 및 기대 효과

a. 결론

랜덤 포레스트 기반 모델은 날씨 조건과 교통사고 발생 간의 관계를 효과적으로 학습했으며, 약 85%의 정확도를 보여준다. 특히 비와 눈 조건에서 사고 위험이 높은 것으로 분석되었으며, 이를 통해 위험도를 사전에 예측 가능

b. 기대효과

1) 교통사고 발생 위험 경고 시스템 도입 가능
2) 정책 수립 시 과학적이고 체계적인 의사결정 지원

3) 교통사고 예방을 통해 시민 안전 증대 및 사회적 비용 절감

c. 머신러닝 모델의 한계

자료의 다양성 부족(2023 서울시), 다른 변수 누락(교통량, 시간대, 도로 환경 등), 과적합(데이터가 적은 경우 문제발생)

6. 느낀점

이번 과제를 하면서 머신러닝을 이용하면 안전에 대한 프로그램을 만들 수 있는 것을 실감하였고 코드를 짜는데 있어 많은 에러가 발생하면서 만들었고 이 과정에서 모델 개선 및 데이터 이해에 큰 도움을 받았던 것 같다. 이것은 예측 모델이라는 것이기 때문에 주의를 주어 피해를 최소화하는 것에 중점을 두었다고 생각이 들었다. 자료를 좀더 찾아서 넣었으면 좋았을 거라는 생각이 있었다. 나중에 시간대, 교통량, 도로환경 등 추가하여 좀더 정밀한 예측 모델을 만들 수 있을 것 같다고 생각한다.