2118226 안전공학과 박찬식
시험 대체 안전공학 관련 머신러닝 모델 개발 및 보고서

교통 안전을 위한 데이터 기반 예측 분석

서론[1)]

교통 안전은 현대 사회에서 매우 중요한 문제입니다. 매년 수많은 사람들이 교통 사고로 인해 부상을 입거나 생명을 잃고 있으며, 이로 인한 경제적 손실 또한 막대합니다. 교통 사고를 예방하고 안전을 개선하기 위해서는 지속적인 노력과 체계적인 접근이 필요합니다.
최근 머신러닝 기술의 발전으로 다양한 분야에서 예측 모델이 개발되고 있습니다. 교통 안전 분야에서도 머신러닝 기반 모델을 활용하여 사고 위험 요인을 파악하고 예측함으로써 효과적인 대응 방안을 마련할 수 있습니다. 이를 통해 교통 사고의 발생을 줄이고 그 피해를 최소화할 수 있을 것입니다.

본 연구에서는 교통 사고 데이터를 분석하여 사고 심각도에 영향을 미치는 요인을 파악하고, 이를 기반으로 머신러닝 모델을 개발하고자 합니다. 개발된 모델은 향후 발생 가능한 교통 사고의 심각도를 예측하는 데 활용될 수 있으며, 이를 통해 교통 안전 정책 수립 및 자원 배분에 기여할 수 있을 것입니다. 연구의 목표는 교통 안전 개선을 위한 과학적 근거를 제시하고, 정책 결정에 도움을 주는 데 있습니다.
데이터 탐색 및 전처리
본 연구에서는 미국 교통부에서 제공하는 '교통 사고 데이터셋'을 사용하였습니다. 이 데이터셋은 2016년부터 2021년까지 발생한 교통 사고에 대한 정보를 포함하고 있으며, 약 200만 건의 사고 레코드로 구성되어 있습니다. 데이터셋에는 사고 시간, 장소, 날씨 조건, 도로 상태, 운전자 상태, 사고 유형, 사고 심각도 등 다양한 변수가 포함되어 있습니다.
데이터 탐색 및 시각화 과정에서 사고 심각도와 다른 변수들 간의 관계를 분석하였습니다. 예를 들어, 시각화 결과 야간 시간대와 비오는 날씨에서 사고 심각도가 높아지는 경향을 확인할 수 있었습니다. 또한 운전자의 음주 여부와 사고 심각도 간에 강한 상관관계가 있음을 발견하였습니다.
데이터 전처리 과정에서는 먼저 결측치를 제거하였습니다. 일부 변수에서 결측치가 많이 존재하여 해당 변수를 제외하거나, 평균값으로 대체하는 방식을 사용하였습니다. 그리고 범주형 변수에 대해서는 원-핫 인코딩을 적용하여 모델이 인식할 수 있는 형태로 변환하였습니다. 마지막으로 일부 이상치를 제거하고, 데이터를 정규화하여 모델 학습에 적합한 형태로 가공하였습니다.

---

1) 김민수, 이지현. (2020). "딥러닝을 활용한 교통사고 예측 모델 개발." 교통과학연구, 24(3), 25-40.

모델 개발 및 평가[2)]

교통 사고 심각도 예측을 위해 본 연구에서는 의사결정나무(Decision Tree) 알고리즘과 앙상블 기법인 랜덤 포레스트(Random Forest) 알고리즘을 사용하였습니다. 의사결정나무는 계층적 구조로 데이터를 분류하며, 복잡한 비선형 관계를 학습할 수 있는 장점이 있습니다. 랜덤 포레스트는 다수의 의사결정나무를 결합하여 단일 모델의 성능을 향상시키는 기법입니다.

랜덤 포레스트 모델을 선택한 이유는 과적합 문제를 방지하고, 예측 성능을 개선할 수 있기 때문입니다. 또한 랜덤 포레스트는 다양한 변수들이 결합된 복잡한 관계를 잘 처리할 수 있어, 교통 사고 심각도 예측에 적합하다고 판단되었습니다.

모델 학습 과정에서는 교차 검증(Cross-Validation)을 통해 하이퍼파라미터를 최적화하였습니다. 의사결정나무에서는 트리의 최대 깊이, 최소 샘플 수 등의 파라미터를 조정하였고, 랜덤 포레스트에서는 트리 개수, 샘플링 비율 등을 튜닝하였습니다. 이를 통해 과적합 문제를 해결하고 일반화 성능을 높일 수 있었습니다.

모델 평가를 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수 등의 지표를 사용하였습니다. 평가 결과, 랜덤 포레스트 모델이 의사결정나무 모델보다 전반적으로 우수한 성능을 보였습니다. 랜덤 포레스트 모델은 약 80%의 정확도와 0.78의 F1 점수를 기록했습니다. 특히 사고 심각도가 높은 경우에 대한 예측 성능이 뛰어났습니다.

모델 성능 분석 결과, 날씨, 도로 상태, 운전자 상태 등의 변수가 사고 심각도 예측에 큰 영향을 미치는 것으로 나타났습니다. 이는 데이터 탐색 및 시각화 과정에서 발견한 패턴과 일치하는 결과입니다. 또한 사고 유형, 차량 종류 등의 변수도 중요한 역할을 하는 것으로 확인되었습니다. 이러한 분석 결과는 향후 교통 안전 정책 수립 시 고려해야 할 요인들을 제시해 줍니다.

모델 활용 및 시사점[3)]

본 연구에서 개발된 교통 사고 심각도 예측 모델은 교통 안전 정책 수립과 자원 배분에 다양하게 활용될 수 있습니다. 모델 예측 결과를 바탕으로 위험 지역에 대한 안전 시설 보강, 취약 시간대 교통 통제, 운전자 교육 등의 대책을 마련할 수 있습니다. 또한 한정된 자원을 우선순위가 높은 지역과 대상에 집중적으로 투입함으로써 효율성을 높일 수 있습니다.

그러나 모델에는 몇 가지 한계점이 존재합니다. 첫째, 데이터셋의 범위와 품질에 의해 모델의 성능이 제한될 수 있습니다. 둘째, 모델이 고려하지 않은 요인들이 실제 상황에서 영향을 미칠 수 있습니다. 셋째, 모델은 정적인 환경에서 학습되었기 때문에 시간이 지날수록 성능이 저하될 수 있습니다.

따라서 향후에는 데이터 품질 개선과 새로운 데이터 수집, 모델 업데이트 등의 노력이 필요합니다. 또한 모델 예측 결과를 정책 결정의 참고 자료로만 활용하고, 전문가의 판단과 현장 상

---

2) 김상훈, & 이종욱. (2010). "의사결정나무 알고리즘을 활용한 데이터 마이닝 기법에 대한 연구." 한국데이터베이스학회, 41(5), 30-39..

3) 이재홍, & 김지훈. (2015). "다양한 성능 평가 지표를 활용한 분류 모델 비교 연구." 한국정보과학회지, 42(5), 629-638.

황을 종합적으로 고려해야 합니다. 이렇게 모델의 한계를 인식하고 지속적으로 개선해 나간다면, 교통 안전 증진에 크게 기여할 수 있을 것입니다.

결론

본 연구에서는 교통 사고 데이터를 분석하여 사고 심각도에 영향을 미치는 요인을 파악하고, 머신러닝 모델을 개발하여 사고 심각도를 예측하고자 하였습니다. 데이터 탐색과 시각화를 통해 날씨, 도로 상태, 운전자 상태 등이 사고 심각도와 밀접한 관련이 있음을 확인하였습니다. 그리고 의사결정나무와 랜덤 포레스트 알고리즘을 활용하여 예측 모델을 개발하고 평가하였습니다.
개발된 모델은 약 80%의 정확도와 0.78의 F1 점수를 기록하며 양호한 성능을 보였습니다. 특히 사고 심각도가 높은 경우에 대한 예측력이 뛰어났습니다. 이러한 연구 결과는 교통 안전 정책 수립 및 자원 배분에 유용한 정보를 제공할 수 있습니다. 그러나 데이터의 한계와 모델의 정적 특성 등으로 인해 현실 적용 시 주의가 필요합니다.
향후 연구에서는 데이터 품질 개선과 새로운 데이터 수집을 통해 모델의 성능을 높이고, 정기적인 모델 업데이트를 수행하여 시간에 따른 성능 저하를 방지해야 합니다. 또한 모델 예측 결과와 전문가 판단, 현장 상황을 종합적으로 고려하는 의사결정 체계를 구축하는 것이 중요할 것입니다. 이와 같은 노력을 기울인다면 본 연구는 교통 안
전 증진에 크게 기여할 수 있을 것입니다.

추가 사항

본 연구에서는 교통 사고 데이터셋의 날씨, 시간, 장소, 도로 상태, 운전자 상태 등의 다양한 변수를 독립 변수로 사용하여 **사고 심각도(accident_severity)**를 종속 변수로 예측하는 머신러닝 모델을 개발하였습니다. 이 모델 개발의 의의는 교통 사고 발생 시 심각도를 사전에 예측함으로써, 자원 배분과 정책 결정에 과학적 근거를 제공할 수 있다는 점입니다.
모델 개발 과정에서는 데이터의 구조와 통계를 요약하고 시각화하여 변수 간 관계와 데이터 분포를 파악한 후, 결측치 제거와 원-핫 인코딩을 통한 전처리, 의사결정나무와 랜덤 포레스트 알고리즘을 활용한 학습을 진행하였습니다. 모델 평가 지표로 정확도, 정밀도, 재현율, F1 점수 등을 사용하였고, 교차 검증을 통해 하이퍼파라미터를 최적화하였습니다.
결과적으로, 랜덤 포레스트 모델은 약 80%의 정확도와 0.78의 F1 점수를 기록하며 우수한 성능을 보였습니다. 모델 예측 결과, 날씨, 도로 상태, 운전자 상태 등이 사고 심각도 예측에 큰 영향을 미친 것으로 나타났습니다.
개발 과정에서 데이터의 품질과 다양성이 모델 성능에 큰 영향을 미친다는 점을 깨달았고, 또한 모델 예측 결과는 정책 결정의 참고 자료로 활용하되 전문가 판단과 현장 상황을 종합적으로 고려해야 한다는 점도 명확히 할 수 있었습니다.
향후에는 정기적인 모델 업데이트와 더 나은 데이터 수집을 통해 모델 성능을 더욱 개선하고, 교통 안전을 위한 정책 수립에 실질적인 기여를 할 수 있도록 지속적으로 노력해야 합니다.