교통대생의 교통사고 예측모델 만들기

이름: 이시현

학번: 2318084

Github:

1. **안전 관련 머신러닝 모델 개발 관련 요약**

   a. 본 프로젝트의 목적은 교통사고 데이터를 기반으로 사고 유형을 예측하는 머신러닝 모델을 개발하는 목적입니다. 사고 유형 예측은 사고 발생 후 빠른 대응 및 예방 전략을 수립하는정보를 제공하고 도움을 줄것 이라고 생각합니다. 이러한 예측을 통해 교통사고를 예방하고, 사고 유형에 맞는 대응을 신속하게 진행할 수 있도록 지원하려는 목표를 가지고 있습니다.

2. **개발 목적**

   a. **머신러닝 모델 활용 대상:** 이 모델은 교통사고 데이터를 바탕으로 사고 유형을 예측하고, 도로공사, 경찰청, 보험사 등이 사용할 것으로 예측 됩니다.

   b. **개발의 의의:** 사고를 줄임과 동시에 사고 예방 또는 교통 안전을 향상시킬것이라고 생각하고 효율적 대응방안 마련, 사회적인 비용을 줄이고 국민의 생명과 재산을 보호하는데 기여합니다.

   c. **데이터의 어떠한 독립 변수를 사용하여 어떠한 종속 변수를 예측하는지:** 교통사고 데이터를 사용하여 사고 유형을 예측합니다. 독립 변수는 사고 건수, 사망자 수, 중상자 수, 경상자 수, 부상 신고자 수와 같은 사고와 관련된 여러 변수들이며, 종속 변수는 '사고 유형 분류'입니다. 이 데이터를 기반으로 각 사고가 어떤 유형에 해당하는지를 예측합니다.

3. **배경지식**

   a. **데이터 관련 사회 문제 설명**

   교통사고는 전 세계적으로 큰 사회적 문제입니다. 사고 발생 시 많은 사람들의 생명과 재산이 피해를 입으며, 정부 및 사회는 사고를 예방하고, 사고 발생 후 신속하게 대응할 수 있는 시스템을 구축하는 것이 중요하다고 생각합니다. 사고 유형 예측은 사고 발생 원인과 특성을 파악하고, 적절한 예방 및 대응 방법을 제시할 수 있게 도와줄 것입니다.

b. **머신러닝 모델 관련 설명 등**

머신러닝은 과거의 데이터를 바탕으로 모델을 학습시키고, 이를 통해 새로운 데이터에 대한 예측을 수행하는 기술이며. 교통사고 데이터를 바탕으로 사고 유형을 예측하는 데 머신러닝 모델을 사용하는 이유는, 데이터에서 숨겨진 패턴을 찾아내고, 이를 기반으로 보다 정확한 예측을 할 수 있기 때문입니다. 특히, 교차 검증, 하이퍼파라미터 튜닝 등을 통해 모델의 성능을 최적화하는 과정은 모델이 실제 환경에서도 신뢰할 수 있는 예측을 할 수 있도록 만들어 줍니다.

4. 개발 내용

a. 데이터에 대한 구체적 설명 및 시각화

사용된 데이터는 교통사고에 관련된 다양한 변수들을 포함한 데이터셋으로, 사고 건수, 사망자 수, 중상자 수, 경상자 수, 부상 신고자 수 등의 수치형 변수와 사고 유형을 나타내는 범주형 변수 ('사고 유형 대분류')가 포함됩니다.

i. 데이터 개수, 데이터 속성 등

2000건 이상의 사고 데이터를 포함하고 있으며, 각 사고에 대한 다양한 속성을 기록하고 있습니다.

사고 건수, 사망자 수, 중상자 수, 경상자 수, 부상 신고자 수 등 여러 변수들이 포함되어 있으며, 이는 사고의 심각도와 특성을 평가하는 데 중요한 요소들입니다.

ii. 데이터 간 상관관계 설명 등

각 변수 간에는 상관관계가 존재합니다. 예를 들어, 사고 건수와 사망자 수는 높은 상관관계를 보일 수 있습니다. 이를 바탕으로, 변수 간의 상관관계를 분석하여 예측 모델을 개선할 수 있습니다.

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

본 모델은 '사고 유형 대분류'를 예측하는 문제를 다룹니다. 사고 건수, 사망자 수, 중상자 수 등의 독립 변수들을 사용하여 사고 유형을 예측하며, 이를 통해 각 사고에 대해 적절한 대응을 예측할 수 있습니다.

i. 독립변수, 종속변수 설정

독립 변수: '사고건수', '사망자수', '중상자수', '경상자수', '부상신고자수'
종속 변수: '사고유형대분류' (범주형 변수)

c. **머신러닝 모델 선정 이유**

RandomForest는 다수의 결정 트리를 결합한 앙상블 모델로, 교차 검증을 통해 다양한 변수들의 상호작용을 학습하며 높은 정확도를 보입니다. 이 모델은 분류 문제에서 일반적으로 좋은 성능을 보여, 사고 유형 예측에 적합합니다.

i. **설명한 데이터를 기반으로 머신러닝 모델 선정 이유 설명**

이 프로젝트에서 사용된 데이터는 교통사고 관련 데이터를 포함하고 있으며, 사고건수, 사망자수, 중상자수, 경상자수, 부상신고자수와 같은 특징 변수들을 바탕으로 사고유형대분류라는 범주형 목표 변수(사고 유형)를 예측하는 문제입니다. 이러한 문제는 분류(Classification)문제로, 목표 변수는 여러 개의 카테고리로 나뉘어 있기 때문에 다중 클래스 분류문제에 해당합니다.

따라서, 다중 클래스 분류 문제에 적합한 머신러닝 모델을 선택해야 했습니다. 이를 위해 다음과 같은 이유로 랜덤 포레스트(Random Forest)모델을 선택했습니다:

ii. **성능 비교를 위한 머신러닝 모델 선정 이유**

성능 비교를 위한 모델을 선정하는 이유는 모델의 성능을 객관적으로 평가하고, 최적의 모델을 선택하기 위해서입니다.

d. **사용할 성능 지표**

본 모델은 주로 'Accuracy' (정확도), 'Confusion Matrix' (혼동 행렬), 'Classification Report' (분류 보고서)를 사용하여 성능을 평가합니다. 또한, 다중 클래스 문제이므로 'ROC Curve' (수신자 조작 특성 곡선)와 AUC (곡선 아래 면적)를 사용하여 모델의 성능을 평가합니다.

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

ii. 성능 지표 선정 이유 등

Accuracy: 전체 예측 중 정확하게 분류된 비율을 측정합니다.

Confusion Matrix: 모델이 예측한 클래스와 실제 클래스를 비교하여, 모델의 오분류 상황을 시각적으로 확인할 수 있습니다.

ROC Curve & AUC: 다중 클래스 문제에서 각 클래스별로 성능을 평가할 수 있는 유용한 도구입니다.

5. 개발 결과

a. 성능 지표에 따른 머신러닝 모델 성능 평가

i. 수치 자료 및 시각화 자료를 사용

1. MAE, RMSE, MSE, Accuracy, 오차행렬 등

```
from sklearn.metrics import mean_absolute_error

mae = mean_absolute_error(y_test, y_pred)
print(f"평균 절대 오차 (MAE): {mae:.4f}")
```

MAE

```
from sklearn.metrics import mean_squared_error
import numpy as np

rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print(f"루트 평균 제곱 오차 (RMSE): {rmse:.4f}")
```

RMSE

```
mse = mean_squared_error(y_test, y_pred)
print(f"평균 제곱 오차 (MSE): {mse:.4f}")
```

MSE

```
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(y_test, y_pred)
print(f"모델 정확도 (Accuracy): {accuracy:.4f}")
```

Accuracy

```
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=data[target].unique(), ytic
plt.title("Confusion Matrix")
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()
```

혼동행렬

2. **KFold 결과**

2-fold 교차 검증을 통해 데이터의 다양성을 반영하며, 모델이 과적합되지 않도록 학습하였습니다. KFold 결과는 모델의 성능을 더욱 신뢰할 수 있게 해줍니다.

```
from sklearn.model_selection import StratifiedKFold, cross_val_score

# StratifiedKFold 교차 검증 설정
kf = StratifiedKFold(n_splits=2, random_state=42, shuffle=True)

# 교차 검증 수행
cross_val_score_result = cross_val_score(model, X_train_scaled, y_train, cv=kf)
print(f"교차 검증 정확도: {cross_val_score_result.mean():.4f}")
```

ii. **다른 머신러닝 모델과 성능 비교**

다른 모델들과 비교한 결과, RandomForestClassifier가 가장 우수한 성능을 보였으며, 모델이 높은 정확도를 기록하였습니다.

b. **머신러닝 모델의 성능 결과에 대한 해석**

모델은 사고 유형을 정확히 예측할 수 있었으며, 특히 Accuracy와 AUC에서 좋은 성과를 보였습니다. 그러나 여전히 일부 데이터에서 오분류가 발생하며, 이는 추가적인 모델 개선이 필요함을 시사합니다.

6. 결론

a. **머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명**

본 프로젝트에서는 교통사고 데이터를 바탕으로 사고 유형을 예측하는 머신러닝 모델을 개발하였습니다. 사고 유형을 정확하게 예측하였으며, 모델의 성능은 교차 검증을 통해 안정적으로 평가되었습니다.

b. **개발 의의 등**

이 모델은 교통사고의 유형을 실시간으로 예측할 수 있는 가능성을 제시하며, 사고 예방 및 대응 전략 수립에 기여할 수 있습니다. 또한, 교통사고의 위험 요소를 사전에 파악하고, 사고 발생 가능성을 낮출 수 있는 중요한 도구로 작용할 것이라고 생각합니다.

c. **머신러닝 모델의 한계**

데이터 불균형: 일부 사고 유형이 과소 대표되는 문제가 있을 수 있습니다. 이를 해결하기 위해 데이터 샘플링 기법을 적용할 수 있습니다.

오분류: 여전히 일부 사고 유형에서 오분류가 발생할 수 있으므로, 추가적인 모델 개선 및 파라미터 튜닝이 필요합니다.