<데이터 마이닝 프로젝트 보고서>

주제 : 자동차의 종에 따라 분류 및 사고의 규모를 보고 차종을 예측하는 모델 만들기

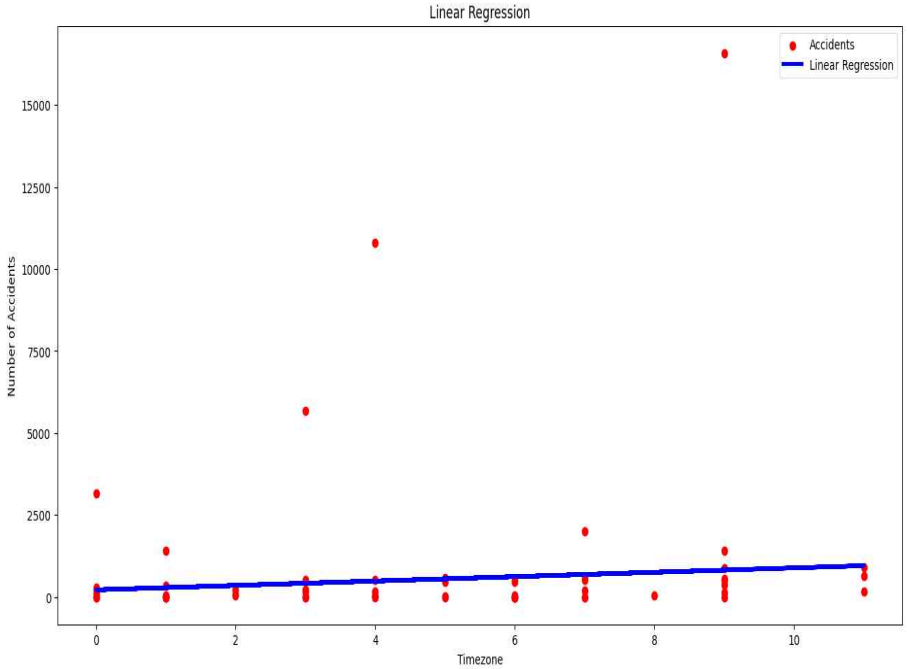
<보고서 내용>

1. 목적

데이터를 분석하여 어떤 시간대에 사고가 가장 발생했는지 파악하고, 사상자의 수에 따라 가해 차량의 차종이 어떤 차량인지를 파악하는 것에 목적을 두고 프로젝트를 진행하였습니다.

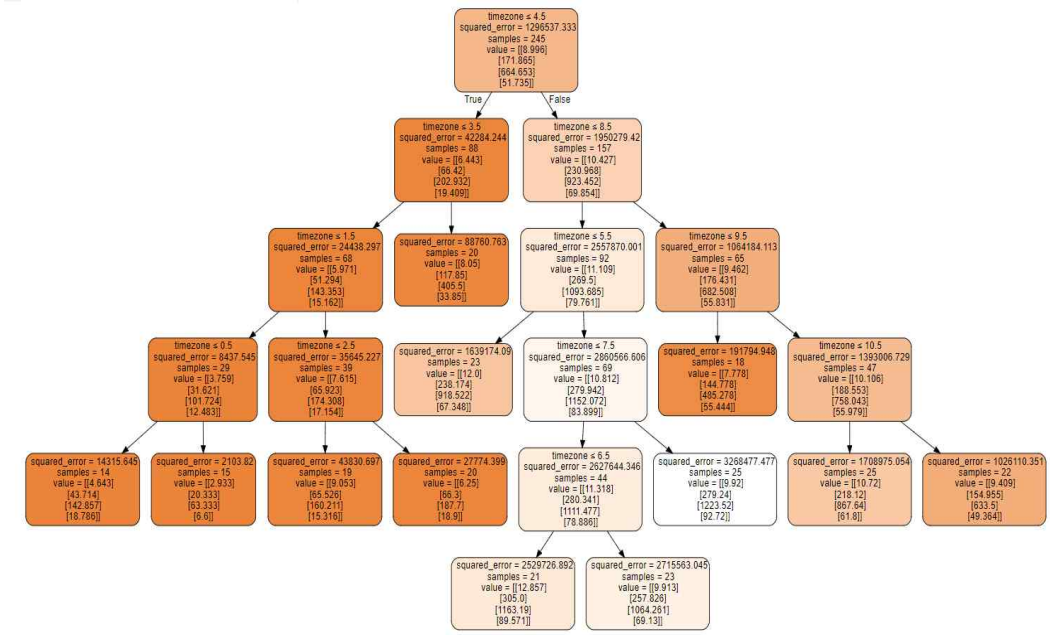
2. 결과

a. 선형회귀



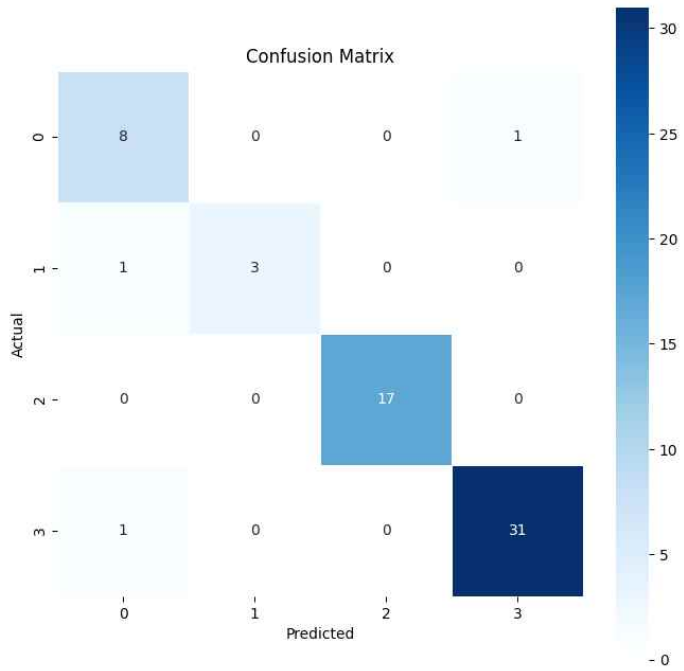
선형회귀를 이용하여, 시간에 따른 사고횟수를 분석하였습니다 시간을 수치화 하였으며 사고횟수가 시간이 지날수록 점점 증가한다는 것을 알 수 있습니다. 특히 Accidents를 보면 9 (18시 ~ 20시)가 가장 큰 사고가 발생 했다는 것을 알 수 있습니다.

b.의사 결정 트리



이번 분석에서는 의사결정트리를 이용하여 시간대(Timezone)를 이용하여 사상자수('number of death', 'number of slanderers', 'number of casualties', 'number of reported injuries')를 분석하였습니다. Mean Squared Error: 2280689.376849166가 나와 오차가 커 이 분석모델은 성능이 좋다고 할 수 없고 예측이 실제와 다르다고 나왔습니다. 고로 의사결정트리를 통한 분석은 안 좋다고 할 수 있습니다.

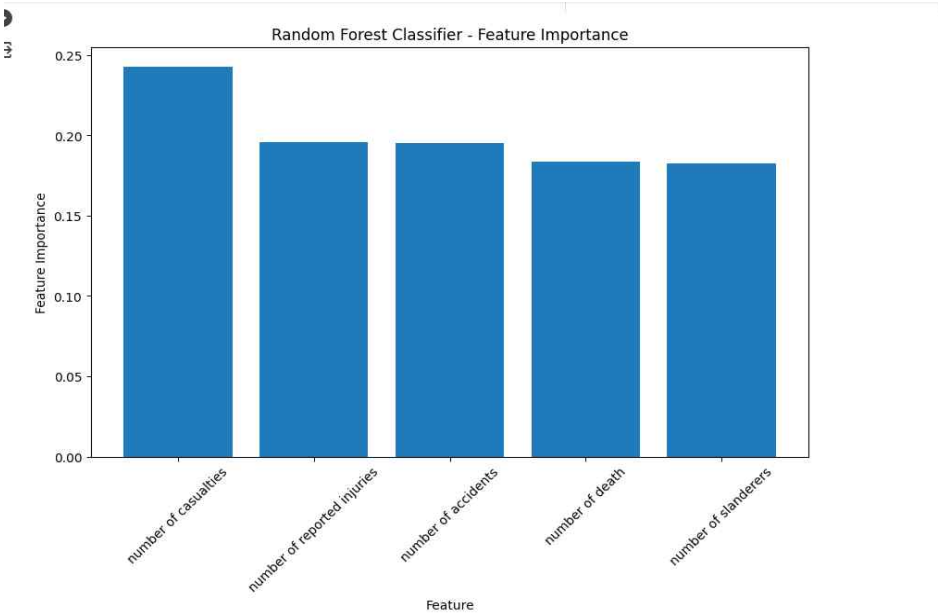
c. 로지스틱 회귀



사고횟수(number of accidents)를 바탕으로 다양한 규모에 대한 사고를 분석 하였습니다 사고횟수를 'No Accident', 'Small', 'Medium', 'Large', 'Very Large'로 나눠서 했고

사상자수를 바탕으로 분석 하였습니다. 모델을 만들고 Accuracy: 0.95161290322258065가 나왔습니다 표를 보면 대각선 값이 크게 나오고 있습니다.

d. 랜덤 포레스트



랜덤 포레스트를 이용하여 어떤 특성이 중요하는지 파악 했습니다. 이 모델의 테스트 성능 평가로 Accuracy: 0.40322580645161290로 나왔으며, 5가지 특성의 이용해서 분석을 했습니다. 그 결과 부상자수(number of casulaties)가 가장 중요한 것으로 가장 높게 나온 걸 알 수 있었습니다.

<결론>
선형회귀, 의사결정트리, 로지스틱회귀, 랜덤 포레스트를 이용하여 다양한 분석을 한 결과를 종합해보자면 시간대가 야심한 밤인 경우에 교통 사고가 더 발생 했으며, 의사결정을 통해 사상자수를 분석해봤지만 MSE의 값이 커서 이러한 모델 분석은 안 좋다는 것을 알 수 있습니다. 또 한 로지스틱 분석을 이용 결과 경미한 사고보단 대형 사고가 자주 발생했다는 것 또한 알 수 있었고 정확도가 높게 나왔습니다. 마지막으로 랜덤 포레스트를 이용한 결과 사상자 분석에서 부상자 수가 가장 높게 나와 교통 사고시에 부상자 수가 많다는 것을 알게 되었습니다.

<사용한 데이터 셋>
https://www.kaggle.com/datasets/kimminky/korea-road-traiffic/

주제 바꾼 이유 : 기존 주제에 있던게 한글이라 도저히 코랩에서 한글을 할수 없어서 주제를 급하게 바꿨습니다. 또한 데이터 csv파일이지만 excel파일로 옮기고 데이터셋에 city ?? bus에서 city bus로 individual 이동형 장치수단 (PM)에서 individual(PM)로 바꿨습니다.