**# 2. 교통사고 위험구간 예측을 위한 파생변수 생성 및 모델링 방법론 제시 (1장 이내)**

**데이터모델링 방안**

**1. 파생변수 빛 데이터셋 관리전략**

[별첨2] 표1, 표2 첨부

표1 : 1차 파생변수 관리전략

표2: 데이터셋 관리전략

**2. 2차 파생변수 및 모델링 전략**

**(1) 2차 파생변수**

- 사고심각도 변수 = f(사고내용, 총사망자 및 부상자 등) : output 점수

- 위치별 사고 변수 = f(도로유형, 도로시설물데이터, 사고빈도 등): output 사고빈도 or 사고확률

- 시간별 사고 변수 = f(주/야, 계절, 날씨 등 ) : output 사고빈도 or 사고확률

- 위험지역 변수 = f(도로유형, 도로시설물데이터, 사고빈도) : output 위치 or 도로유형재범주화

기타변수

- 보행자사고다발지역

- 어린이사고다발지역

- 차대차사고다발지역

**(2) 모델링전략**

사고위험지역 = f(사고심각도 점수, 사고빈도점수, 위험지역점수)

**모델링 방법론**

- test set, training set

- 인공신경망, KNN, 등…

모델 분류하고 ..공부한 표 넣어보기

- input 변수 set을 여러 개 조합하여 예측력이 가장 높은 모델을 선정

-

**데이터 전처리 방안**

1) 결측률이 높은 변수를 제거

2) 파생변수 생성

교통사고 데이터

- 발생일시 변수에서 “날짜”(YYYY-MM-DD), “시간”(00시) 변수 생성

- 시간 변수에서 “주/야” 변수 생성: 주간(06:01~18:00), 야간(18:01~06:00)으로 재범주화

- “총 부상자 및 사상자” 변수 생성: 사망자수 + 중상자수 + 경상자수 + 부상신고자수

- 사고유형 변수에서 “사고유형1” 변수 생성:

3) 오분류된 범주 정리, 재범주화

모델링 방안

