1. **The title of your project**

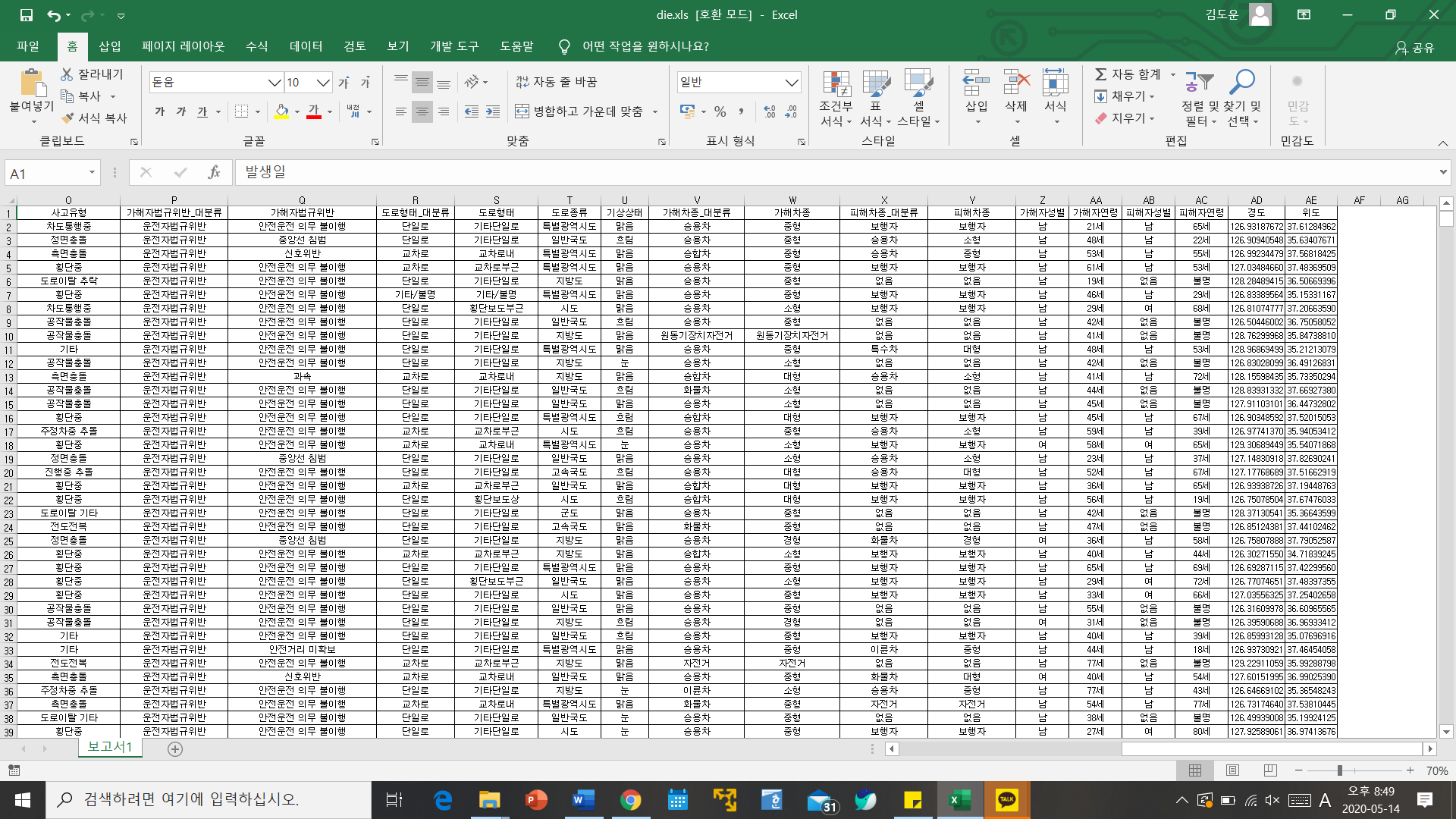
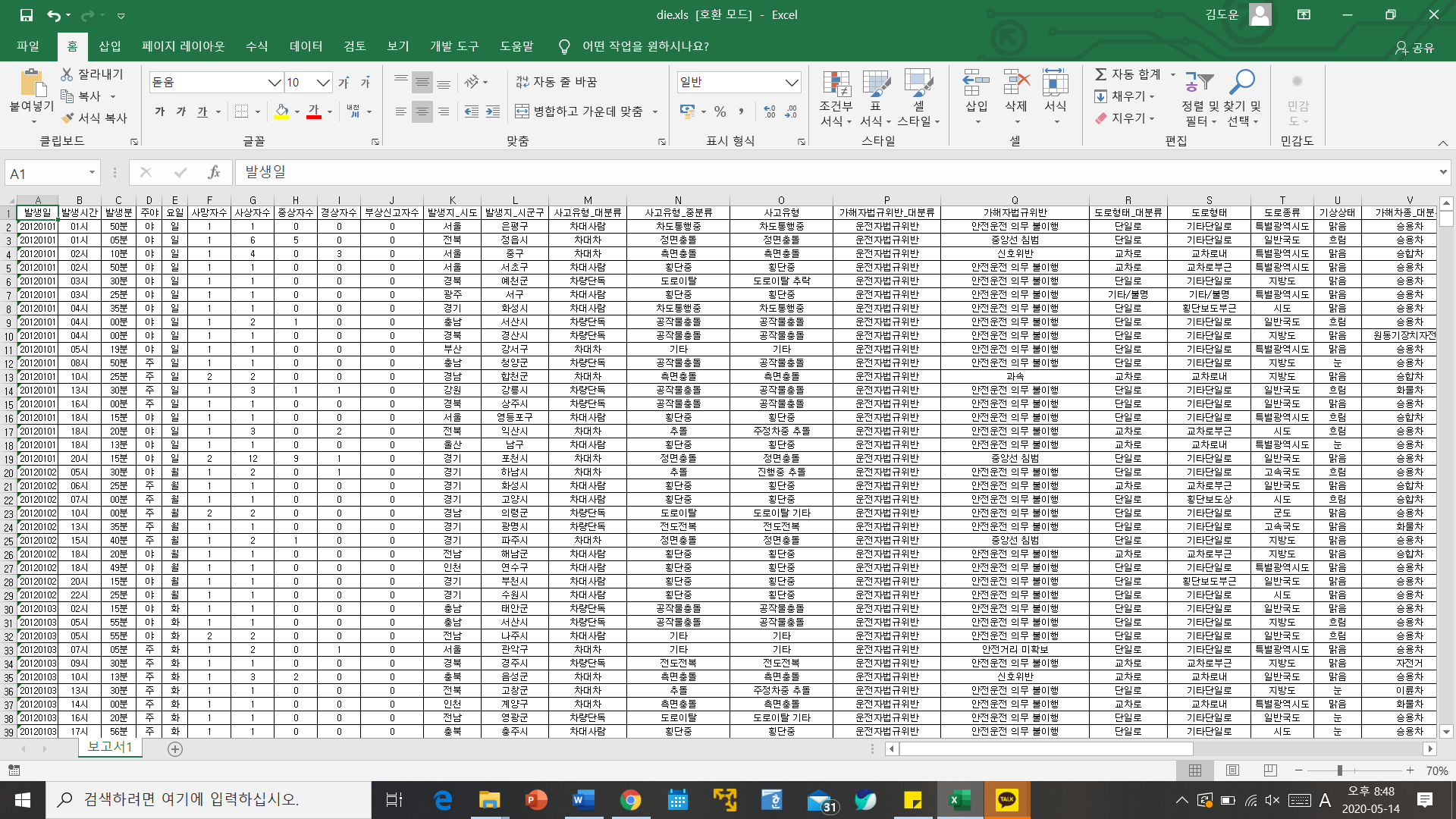
* 미안하다 예측하는거 보여줄려고 어그로 끌었다. 오늘 위험도 실화냐?

1. **Names of your group members**

* 어서오십쇼HUMAN
* 허윤행, 김도운, 김용현, 유성민

1. **What kind of data we will use?**

* [TAAS 교통사고 분석시스템](http://taas.koroad.or.kr/)에서 실제로 발생한 교통사고의 개별정보가 담긴 data를 사용할 예정이며, 해당 dataset에는 ‘발생일시’, ‘발생 요일’, 사고내용’, ‘사고 유형’, ‘가해자법규위반 사항’, ‘노면 상태’, ‘기상 상태’, ‘도로 형태’, ‘피/가해자 성별’, ‘피/가해자 나이’, ‘위도’, ‘경도’, ‘발생지역’ 등(대략 25여가지의 feature 포함)의 다양한 feature들이 있으며, data sample수 또한 5만여 가지로 충분하다. Feature들 중에서는 기상상태, 도로 형태, 노면 상태, 발생 시간대 등의 feature들을 사용해볼 수 있을 것으로 예상된다.



(예시 데이터)

1. **What kind of problem you are solving?**

* **** 각 사고 일자와 요일, 사고 유형, 노면 상태, 기상 상태 등에 따라 사고의 위험도는 어떨까 에 대한 문제를 풀려고 한다. 따라서 이번 프로젝트에서 가장 무게를 두고 있는 최종 주제는 사고 내용을 통한 classification 모델을 통해서 사고 내용을 class들로 나누어서 각 정도를 퍼센트로 나타내어 사고 위험도를 예측해주는 프로그램을 개발할 생각이다. 예를 들면 부상신고, 경상사고, 중상사고, 사망사고로 각 사고 내용을 class 0~3까지로 나타내어 classification문제로 접근하여 분류 및 예측을 하고, 각 예측된 class들에 따라 class 0이면 0~25%, class 1이면 25~50%, class 2이면 50~75%, class 3이면 75%~100%로 퍼센트로 나타내는 방법이다.

1. **How you would solve your problem (main approaches, candidate algorithms)**

* 우선 데이터 자체가 다 categorical하므로 이러한 categorical한 변수들을 numeric하게 바꾸는 전 처리 과정이 필요할 것으로 생각된다. 예를 들면 요일 같은 변수들은 1부터 7로 나타내고, 사고 유형도 마찬가지로 종류에 따른 숫자로 나타내도록 하는 전 처리 과정이 필요할 것이다. 또한 feature들이 다양하므로 변수들 간의 correlation coefficient를 측정하여 상관관계 분석하고 필요한 feature들을 선택하거나 forward selection과 같은 변수 선택 방식을 이용하여 최종 모델에 사용할 변수들을 선택하는 것이 괜찮아 보인다.
* 이 dataset을 활용하여 하고자 하는 것이 분류 모델이므로 classification들로 할 수 있는 모델 중 bagging 기법과 random forest 기법 사용 가능하다고 생각된다. 그 이유는 데이터의 샘플 수도 적지 않고 사용해볼 feature도 많기 때문에 최대한 활용할 수 있는 모델이 bagging과 random forest라고 생각되기 때문이다. 이전의 과제를 통해서 위와 같은 모델들을 사용했을 때 정확도가 높다고 생각되기 때문에 이러한 모델들을 후보로 생각하였으며 추가적으로 다양한 모델들을 생성해서 정확도를 비교함으로써 어떤 모델이 가장 적합한지 선택할 수도 있다.

1. **How you will evaluate your solution**

* 먼저 해당 데이터셋을 90%의 train set과 10%의 test set으로 분리한 뒤, training set을 cross validation 방식과 Grid search 방식으로 적절한 hyper parameter들을 찾은 뒤, 최적의 hyper parameter로 설정된 모델에 test set을 사용하여 최종 점검을 해볼 예정이다. 또한 precision, recall, f1 score들의 점수를 계산해 얼만큼 잘 사고의 정도를 분류하였는지 확인해볼 수 있다.