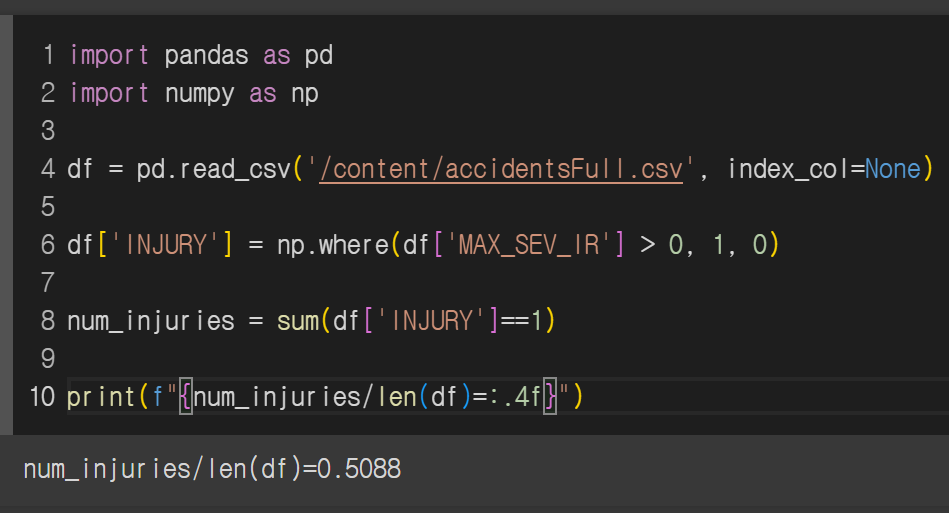
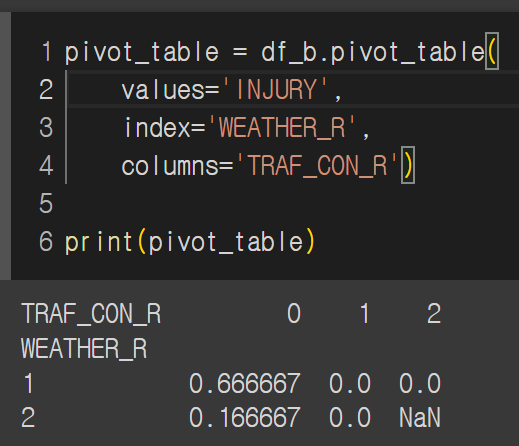
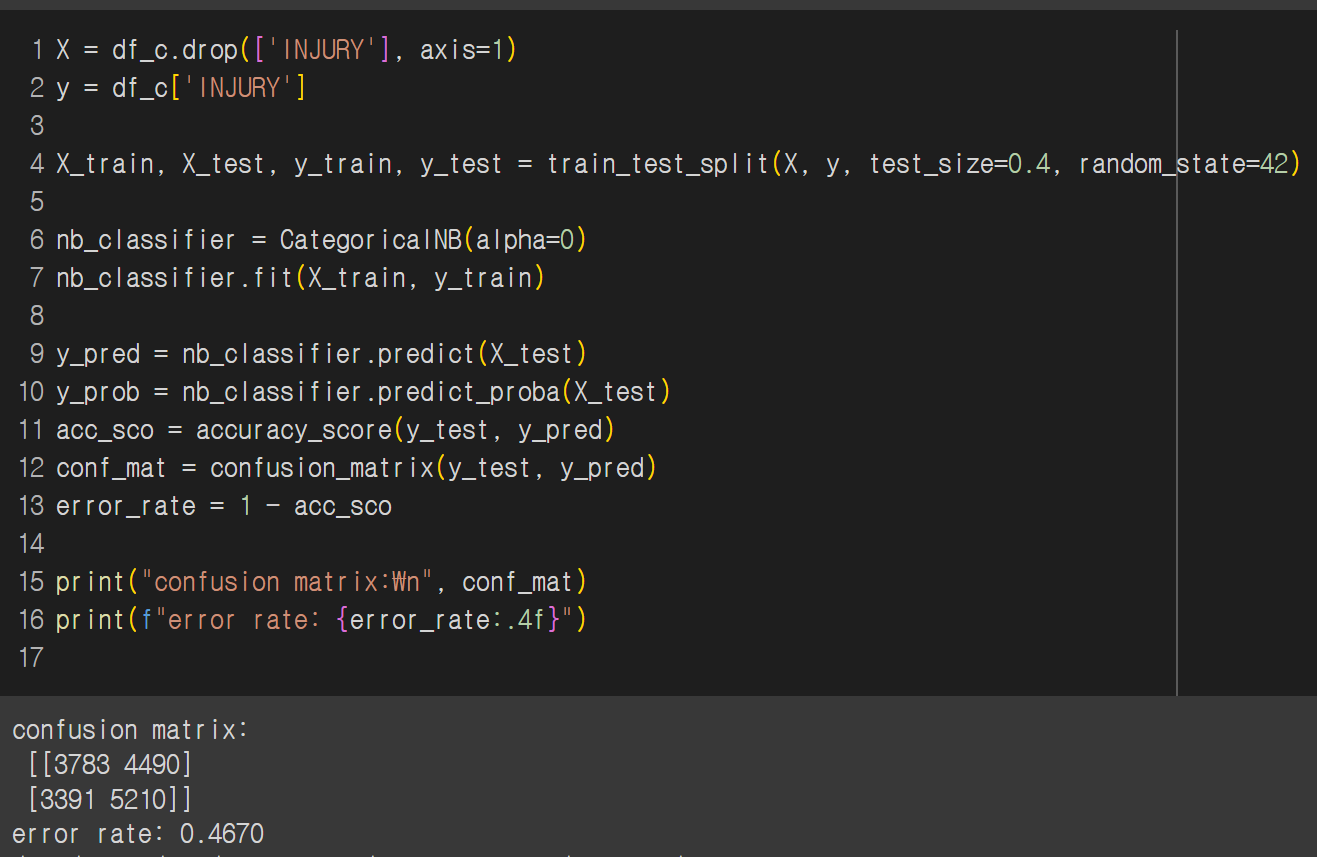
1. 이 데이터셋의 정보를 사용하여, 막 보고된 사고에 추가 정보가 없다면 예측은 무엇이 되어야 하는가? (INJURY = Yes 또는 No?) 그 이유는 무엇인가?  
     
   추가 정보가 없다면 사고시 부상을 당한 경우가 더 많으므로 사고가 보고되었다면 부상일 확률이 높아 부상으로 예측해야 한다  
   
2. 이 데이터셋의 처음 12개 레코드를 선택하여 응답값(INJURY)과 2개의 예측 변수 WEATHER\_R과 TRAF\_CON\_R만 고려하자.  
   텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

   자동 생성된 설명
   1. 12개의 레코드들에 대해 두 예측 변수의 함수로서 INJURY를 검토하는 피벗 테이블을 작성하시오. 피벗 테이블에서 이 세 변수를 모두 행/열로 사용하시오.  
        
      피벗테이블은 index와 columns의 조건에서 values가 true일 확률을 보여준다. 이때, weather=2, traf=2의 조건에 대한 정보는 존재하지 않으므로 NaN으로 표기된다.
   2. 예측 변수들의 가능한 조합 6가지에 대해, 부상이 있을(INJURY = Yes) 정확한 베이즈 조건부 확률(Exact Bayes conditional probability)을 계산하시오.
   3. 이 확률값들과 컷오프 값 0.5를 사용해 이 12개의 사고를 분류하시오.
   4. WEATHER\_R = 1과 TRAF\_CON\_R = 1이 주어졌을 때, 부상이 있을 나이브 베이즈 조건부 확률을 손으로 계산하시오.
   5. scikit-learn을 사용해 2개의 예측 변수를 갖는 12개의 레코드들에 대해 나이브 베이즈 분류기를 실행하시오. 12개 레코드 모두에 대해 확률값과 분류 결과를 구해서 모델의 출력값을 살펴보시오. 이 출력값을 정확한 베이즈 분류 결과와 비교하시오. 분류 결과가 동일한가? 관측치들의 랭킹(=순위)은 동일한가?  
        
      텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

      자동 생성된 설명  
      텍스트, 스크린샷, 메뉴이(가) 표시된 사진

      자동 생성된 설명  
      나이브 베이즈 분류기를 사용하더라도 exact 베이즈 분류기와 분류 결과와 관측치는 동일하나, 확률은 상이한 결과를 보인다.
3. 이제 전체 데이터셋으로 돌아가보자. 데이터를 학습 데이터셋(60%)과 검증 데이터셋(40%)으로 분할하시오.
   1. 예측하는 시점에 사고 자체에 대한 정보나 초기 보고서가 가용하지 않다고 가정할 때(위치 특성과 기상 조건 등만 주어짐), 어떤 예측 변수들을 분석에 포함시킬 수 있는가? (파일의 데이터 명세 페이지를 참조하시오.)  
        
      <https://www.leonshpaner.com/teaching/post/data_science_for_everyone/case_study/>

**\* 시간 정보: HOUR\_I\_R(시간), WKDY\_I\_R(요일), LGTCON\_I\_R(빛 조건)  
\* 위치 정보: ALIGN\_I(도로 형태), WRK\_ZONE(공사 지점), INT\_HWY(고속도로), RELJCT\_I\_R(교차로 사고), REL\_RWY\_R(도로 위 사고 여부), PROFIL\_I\_R(도로 곡면도), SPD\_LIM(제한속도), TRAF\_CON\_R(신호통제장치), TRAF\_WAY(도로 종류)  
\* 기상 정보: SUR\_CON(도로표면상태), WEATHER\_R(날씨)**

* 1. 관련된 예측 변수들(과 결과 변수에 해당하는 INJURY)로 구성된 완전한 학습 데이터셋을 이용해 나이브 베이즈 분류기를 실행하시오. 모든 예측 변수가 범주형이라는 점에 주목하고, 정오 행렬을 나타내시오.
  2. 검증 데이터셋에 대한 전체 오차는 얼마인가?  
      = 0.4670  
     
  3. (검증 데이터셋을 사용한) 나이브 규칙과 비교한 상대적인 퍼센트 향상 정도는 얼마인가?  
     **어떤 것이랑 비교하라는 것인지 문제 이해가 안되네요…**
  4. 피벗 테이블들의 조건부 확률들을 검토하시오. P(INJURY=No ∣ SPD\_LIM=5)에 대해 확률값이 0이 되는 이유는 무엇인가?  
       
     텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

     자동 생성된 설명 **해당 확률값이 0이 되지 않네요….ㅠㅉ**