**[열 번째 과제]**

**IT응용시스템공학과**

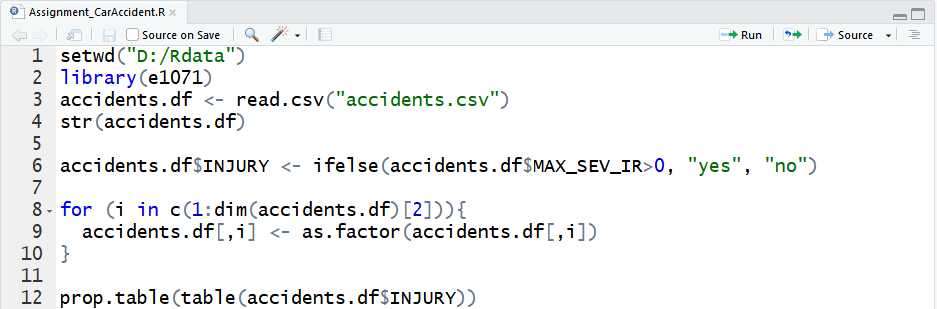
**1494053**

**김희택**

**8.2 자동차 사고 Accidents.csv 파일에는 세 레벨의 부상 (NO INJURY, INJURY, FATALITY) 중의 하나와 연관된 2001년 미국의 실제 자동차 사고 42,183건에 대한 정보가 수록되어 있다. 각 사고에 대해서 요일, 기상조건, 도로종류와 같은 추가적인 정보가 기록되어 있다. 어떤 회사가 초기 보고서와 이 시스템의 연관된 데이터(그 중 일부는 GPS-지원 보고에 의존함)에 근거하여 사고의 심각성을 신속하게 분류하는 시스템을 개발하고자 한다.**

**여기서 우리의 목표는 보고된 사고가 부상이 동반될지 (MAX\_SEV\_IR=1이나 2), 없을지(MAX\_SEV\_IR=0)를 예측하는 것이다. 이러한 목적을 위해서 MAX\_SEV\_IR=1이나 2면 “Yes”를 취하고 그렇지 않으면 “No”를 취하는 INJURY라는 가변수를 생성하시오.**

1. **이 데이터세트의 정보를 사용하여 사고가 막 보고되는 데 추가적인 정보가 없다면, 예측은 무엇이 되어야 하나? (INJURY=Yes 또는 No?) 그 이유는?**

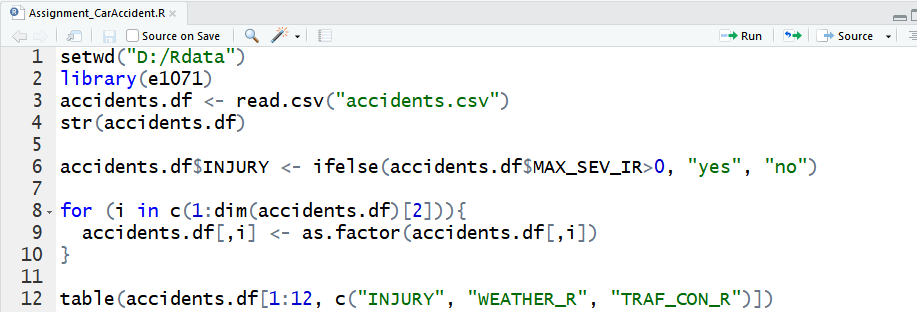


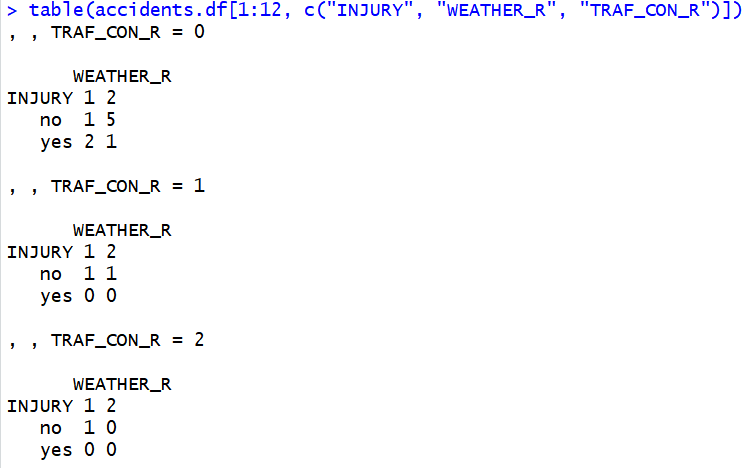
* MAX\_SEV\_IR이 ‘1, 2’일 경우 “Yes”를 취하고, 그렇지 않으면 “No”를 취하는 가변수 INJURY를 생성하였다. 또한, 모든 변수가 범주형이 되도록 factor를 이용하였다.



* 위 결과는 나이브 규칙을 적용한 것이다.
* 추가적인 정보가 없으므로 no와 yes의 수치를 확인했을 때, **INJURY는 더 높은 값인 yes로 예측이 될 것이다.**

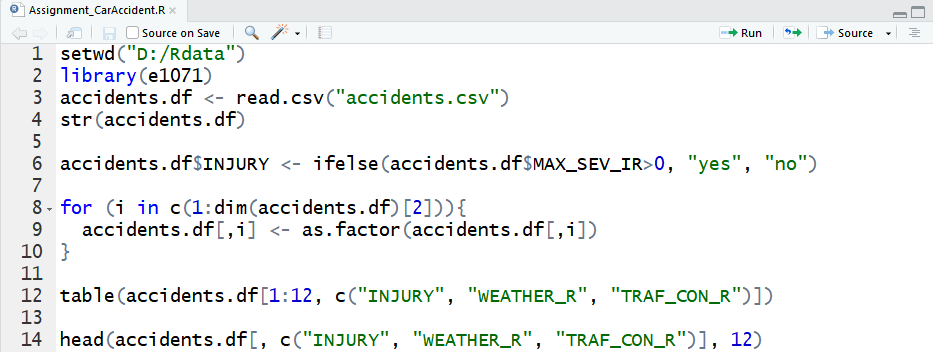
1. **이 데이터세트의 처음 12개 레코드를 선택하여 응답(INJURY)과 두 개의 예측변수 WEATHER\_R과 TRAF\_CON\_R만을 고려하시오.**
2. **이 12개의 레코드들에 대해서 두 예측변수들의 함수로서 INJURY를 행변수로 사용하는 table() 함수를 이용한 2차원 혹은 3차원 테이블을 작성하시오.**

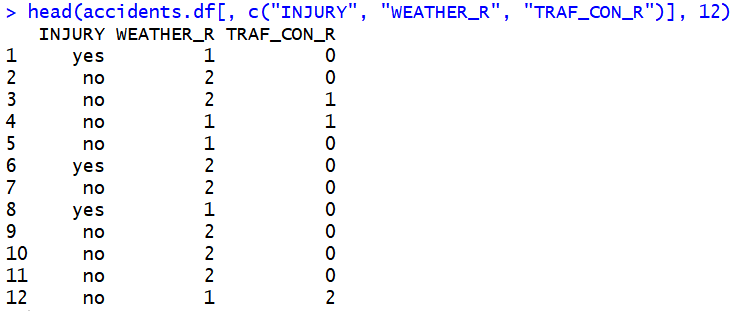




* 위 결과는, 세 개의 변수가 설정되는 경우에 첫 번째 변수는 행이므로 조건에 맞도록 INJURY를 행변수로 사용하고 두 번째 변수(WEATHER\_R)는 열인 교차표를 세 번째 변수(TRAF\_CON\_R)의 각각의 값에 대하여 출력하는 3차원 테이블이다.

1. **예측변수들의 여섯 개 가능한 조합이 주어졌을 때 부상이 있을 (INJURY=Yes) 정확한 베이즈 조건부 확률을 계산하시오.**





* 먼저, head() 함수를 이용하여 데이터를 출력하였다.

P(Yes | WEATHER\_R=1, TRAF\_CON\_R=0) = 2/3 = 약 0.67  
P(Yes | WEATHER\_R=1, TRAF\_CON\_R=1) = 0/1 = 0 ⬄ P(No | WEATHER\_R=1, TRAF\_CON\_R=2) = 1  
P(Yes | WEATHER\_R=1, TRAF\_CON\_R=2) = 0/1 = 0   
P(Yes | WEATHER\_R=2, TRAF\_CON\_R=0) = 1/6 = 약 0.167  
P(Yes | WEATHER\_R=2, TRAF\_CON\_R=1) = 0/1 = 0  
P(Yes | WEATHER\_R=2, TRAF\_CON\_R=2) = 0 (조건 자체가 없다.)

* 위 결과는, 예측변수 WEATHER\_R와 TRAF\_CON\_R로 가능한 여섯 개 조합에서 부상이 있을 (INJURY=Yes) 정확한 베이즈 조건부 확률을 계산 값이다.

1. **이 확률값들과 컷오프 값 0.5를 사용하여 3개의 사고를 분류하시오.**

P(Yes | WEATHER\_R=1, TRAF\_CON\_R=0) = 2/3 = 약 0.67  
P(Yes | WEATHER\_R=2, TRAF\_CON\_R=0) = 1/6 = 약 0.167 P(Yes | WEATHER\_R=2, TRAF\_CON\_R=1) = 0/1 = 0

|  |  |  |
| --- | --- | --- |
| **Cutoff = 0.5** | **Predicted Class** | |
| **Actual Class** | **Yes** | **No** |
| **Yes** | 1 | 2 |
| **No** | 0 | 0 |

* 위 결과는, 임의의 3개의 사고의 조합으로 컷오프 값 0.5 일 때의 분류이다.

1. **WEATHER\_R=1과 TRAF\_CON\_R=1이 주어졌을 때 부상이 있을 나이브 베이즈 조건부 확률을 수작업으로 계산하시오. (조건부 독립)**

* P(Yes | WEATHER\_R=1, TRAF\_CON\_R=1) =

**=**

\* \* ) =

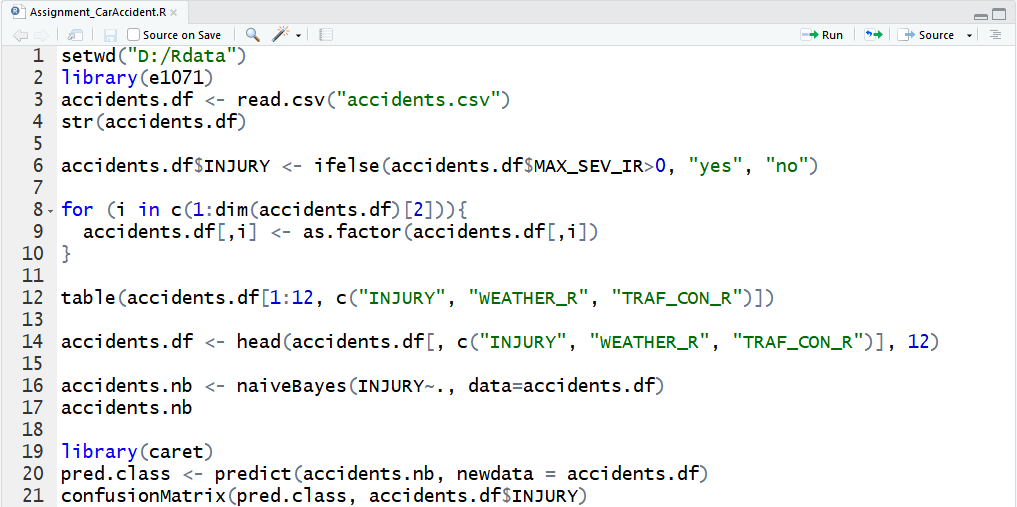
* **(2/3) \* (0/3) \* (3/12)**

**=**

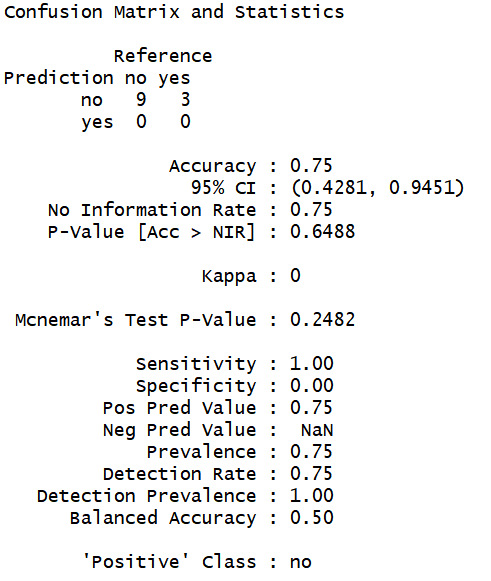
\* \* ) =

* **(3/9) \* (2/9) \* (9/12)**
* 따라서, 분자가 0이 되므로 WEATHER\_R=1과 TRAF\_CON\_R=1이 주어졌을 때 부상이 있을 나이브 베이즈 조건부 확률은 ‘0’이다.

1. **R을 사용하여 12개 레코드에 대해서 나이브베이즈 분류기를 돌려 결과에 대한 정오분류표를 작성하시오. (위에서 주어진 예측변수 2개 사용)**

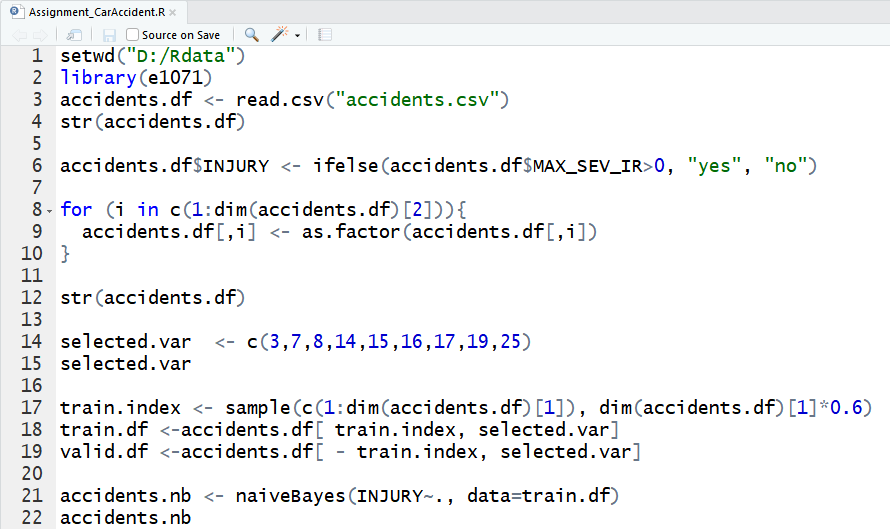


* 12개의 레코드에 대해 나이브베이즈 분류기를 돌려, accidents.nb 결과에 대한 정오분류표를 작성하였다.

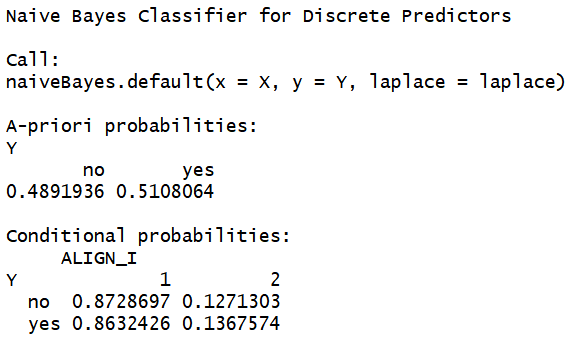


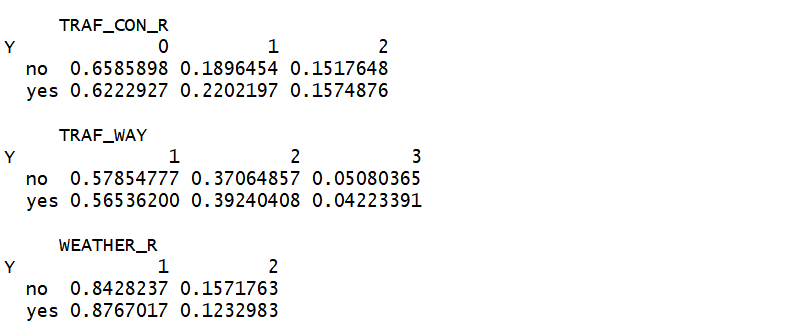
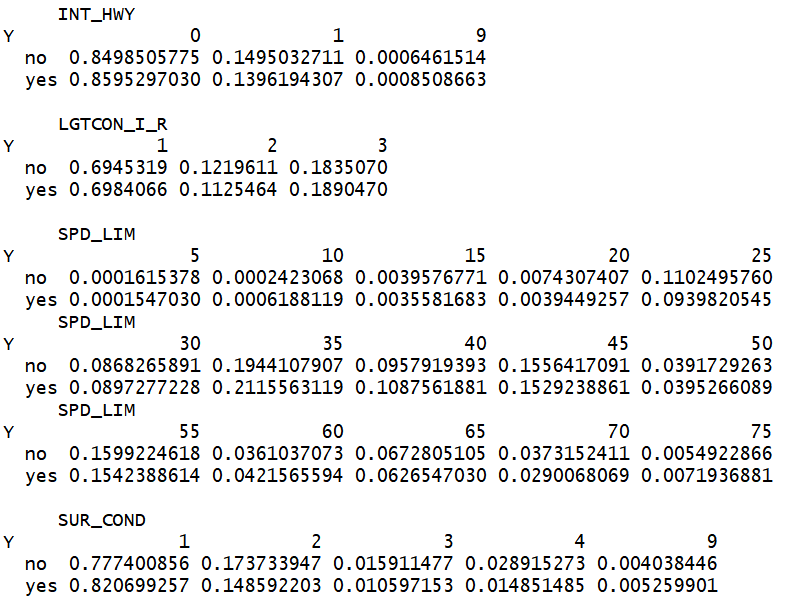
* 위 결과가 정오분류표이다.

1. **이제 전체 데이터세트로 돌아가 보자. 데이터를 학습(60%)과 검증(40%) 세트로 분할하시오.**
2. **관련된 예측변수들(과 응답으로서 INJURY)을 갖는 완전한 학습 세트를 사용하여 나이브 베이즈 분류기를 실행하시오. 모든 예측변수들은 범주형임에 주목하고, 정오행렬을 보이시오.**

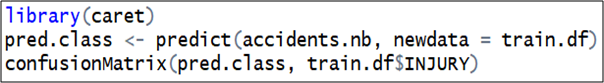


* 먼저, 위ⅰ문제에 따라 위치 특성과 기상조건 등 이와 관련된 변수들을 추출하여 train.df와 valid.df를 생성하였다. 이후 관련된 예측변수들과 완전한 학습 세트를 사용하여 나이브 베이즈 분류기를 실행하였다.

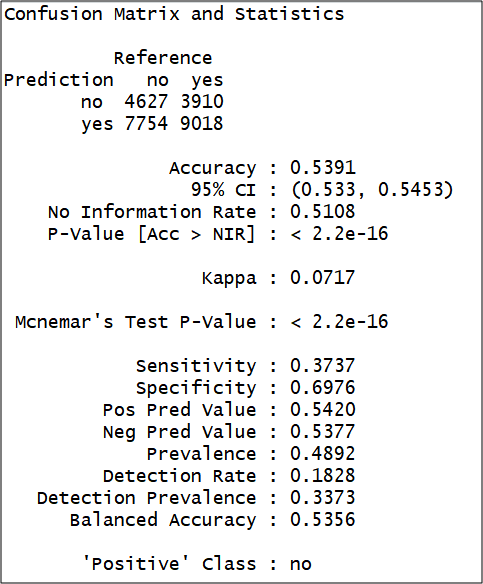




* 위 결과는, 나이브 베이즈 분류기를 실행한 것이다.

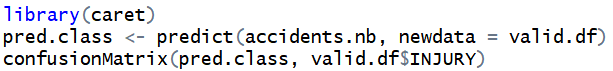


* 학습 데이터세트의 정오행렬을 생성하였다.

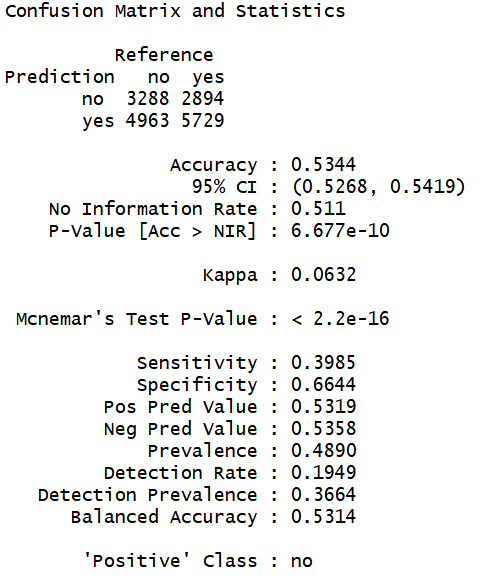


* 위 결과는, 정오행렬을 구한 것이다.

1. **검증 데이터세트에 대한 전체 오차는 얼마인가?**



* 검증 데이터세트의 정오행렬을 생성하였다.



* 전체 오차는 ‘1-Accuracy’ 이므로, 위 결과를 참고하여 1 - 0.534 = **0.466** 인 것을 알 수 있다.
* train.df / valid.df 가 Accuracy가 비슷하므로 과적합 되었다고 볼 수 없다.