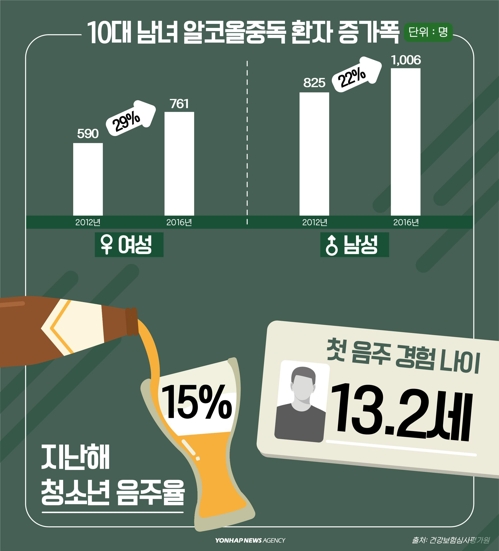
청소년의 음주 수준을 예측하기 위한 데이터 분석

1. 이름(생년월일) : 민두홍(19910906)

2. 지원서에 작성한 Edwith ID : bs11310@naver.com

3. 팀 정보 : 민두홍(19910906) / 안소정(19950619) / 장유진(19950921)

**분석 배경**

호주를 비롯한 많은 국가에서는 음주 허용 연령이 만 18세 이상이다. 쿠바와 같이 아예 음주 연령을 규정하지 않는 나라도 있으며, 세계적인 평균 음주 허용 연령은 15.9세라고 한다. 우리나라의 경우 만 19세 이상부터 주류 구매가 허용되지만 청소년기에 음주를 경험하는 비율이 점점 증가하고 있는 추세다.

지나친 음주는 누구에게나 안 좋은 영향을 미치지만 청소년에게는 특히 치명적이다. 청소년은 성인보다 알코올에 의한 조직파괴가 심각하고 발육부진과 뇌의 발달 및 정신장애 등을 일으키며 건강한 성장을 방해한다. 게다가 청소년기에는 뇌의 가변성이 높아 자극에 쉽게 반응하기에 알코올중독 가능성도 크다. 많은 국가들이 청소년기 음주의 위험성을 인식하고 이에 대한 대책 마련하고 있다. [[1]](#footnote-1)

포르투갈 또한 원래 음주 허용 연령이 16세였지만 2013년부터는 도수가 높은 증류주는 18세 이상으로 규제를 강화했다. 하지만 무조건 음주를 금지하는 것보다는 교육과 청소년들이 음주를 줄일 수 있는 환경을 조성하는 것이 더욱 중요하다. 그러기 위해서는 여러 연구에서 청소년 음주의 주된 동기라고 언급하는 교우관계, 가족영향, 그리고 스트레스에 따른 청소년의 음주 수준을 파악해 볼 필요성을 느꼈다. 그러면서 우리는 스트레스 요인 중 학업에 다른 스트레스가 클 것으로 판단하고 학업과 관련된 요인들을 통해 청소년 음주를 판별할 수 있는지 호기심이 생겼다. 따라서 우리는 지금부터 포르투갈의 청소년들의 성별, 학업, 사회적 요인들에 대한 데이터를 바탕으로 청소년의 음주 수준을 예측해 보고자 한다.

**분석에는 UCI에서 제공하는 “Student Alcohol Consumption”의 데이터를 활용하였다.**

* 데이터 출처 : UC Irvine Machine Learning Repository
* 포르투갈 중학교에서 수학과 모국어를 수강하는 학생들의 성별, 학업, 사회적 요인, 음주 수준에 대한 데이터
* student-mat.csv (Math course)과 student-por.csv (Portuguese language course) 두 가지 데이터셋으로 구분되어 있음

**변수 이름과 설명**

본격적으로 분석을 시작하기에 앞서서 데이터에 포함된 변수들을 유형에 따라 구분해 보았고 각각에 대한 설명을 통해 이해하기 쉽도록 해 보았다.

기본 정보 변수(8)

1. School – 학생의 소속 학교(binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2. sex – 학생의 성별(binary: 'F' - female or 'M' - male)

3. age - 청소년기(15세~22세)의 학생들을 대상을 대상으로 함

4. address – 사는 지역이 도시인지 시골인지 보여줌(binary: 'U' - urban or 'R' - rural)

5. internet – 인터넷 사용 여부(binary: yes or no)

6. romantic – 이성 교제 여부(binary: yes or no)

7. health – 건강 상태(numeric: from 1 - very bad to 5 - very good)

8. goout – 친구와 함께 나가는 횟수(numeric: from 1 - very low to 5 - very high)

가족 변수(6)

9. famsize – 가족 규모(binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

10. Pstatus – 부모가 함께 사는지 여부(binary: 'T' - living together or 'A' - apart)

11. Mjob – 어머니의 직업(nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

12. Fjob - 아버지의 직업(nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')

13. guardian – 학생의 보호자(nominal: 'mother', 'father' or 'other')

14. famrel – 가족관계의 좋고 나쁨(numeric: from 1 - very bad to 5 - excellent)

학습환경 변수(12)

15. Medu – 어머니의 교육수준(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

16. Fedu – 아버지의 교육수준(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

17. nursery – 유치원을 다녔는지 여부(binary: yes or no)

18. higher – 고등교육 진학 계획(binary: yes or no)

19. reason – 학교를 선택한 이유(nominal: close to 'home', school 'reputation', 'course' preference or 'other')

20. traveltime – 통학하는데 걸리는 시간(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

21. studytime – 주당 공부시간(numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

22. famsup – 가족의 학업지원 여부(binary: yes or no)

23. schoolsup – 학교의 추가적인 교육지원 여부(binary: yes or no)

24. paid – 학과 과목(수학 또는 국어)내의 특별 유급 수업 여부(binary: yes or no)

25. activities – 방과 후 활동 여부(binary: yes or no)

26. freetime – 자유시간(numeric: from 1 - very low to 5 - very high).

학업성취 변수(5)

27. absences – 결석 수(numeric: from 0 to 93).

28. failures – 과락 (numeric: n if 1<=n<3, else 4).

29. G1 - first period grade (numeric: from 0 to 20)

30. G2 - second period grade (numeric: from 0 to 20)

31. G3 - final grade (numeric: from 0 to 20)

음주 변수 (2)

32. Dalc – 평일 알코올 섭취 수준 (numeric: from 1 - very low to 5 - very high)

33. Walc – 주말 알코올 섭취 수준 (numeric: from 1 - very low to 5 - very high)

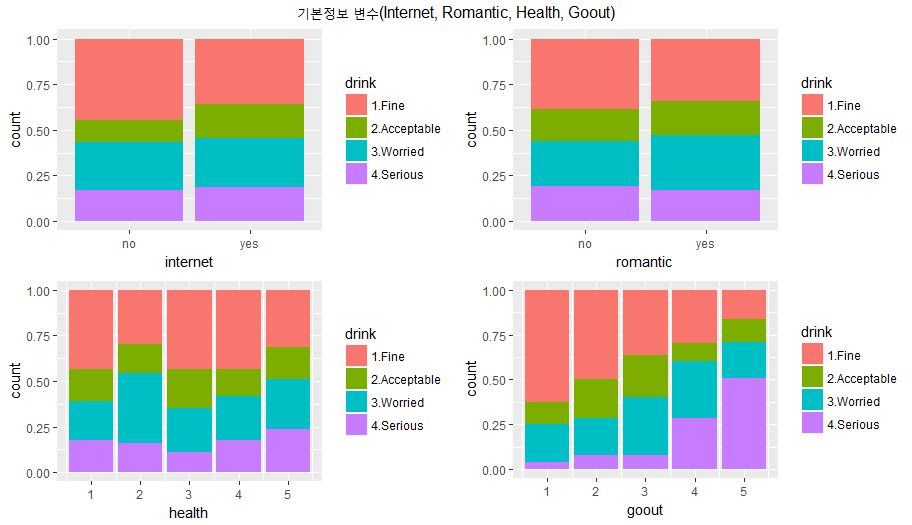
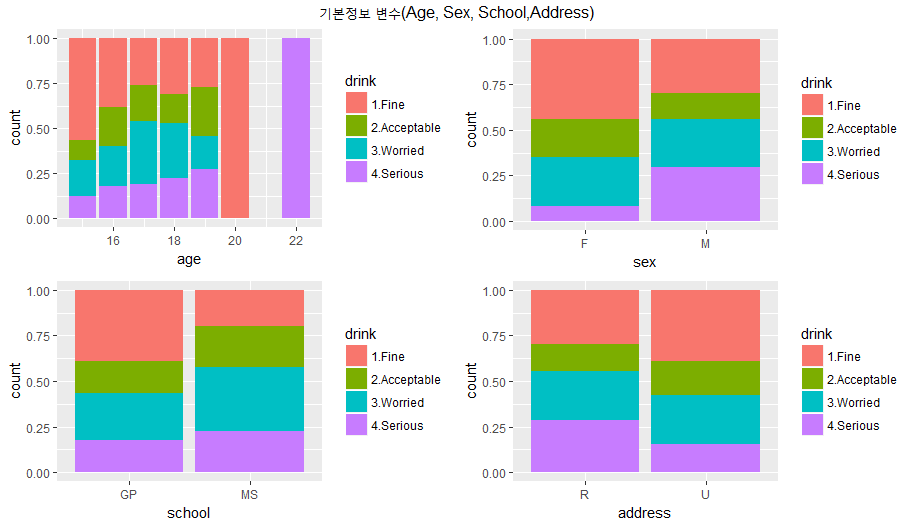
**데이터 전처리**

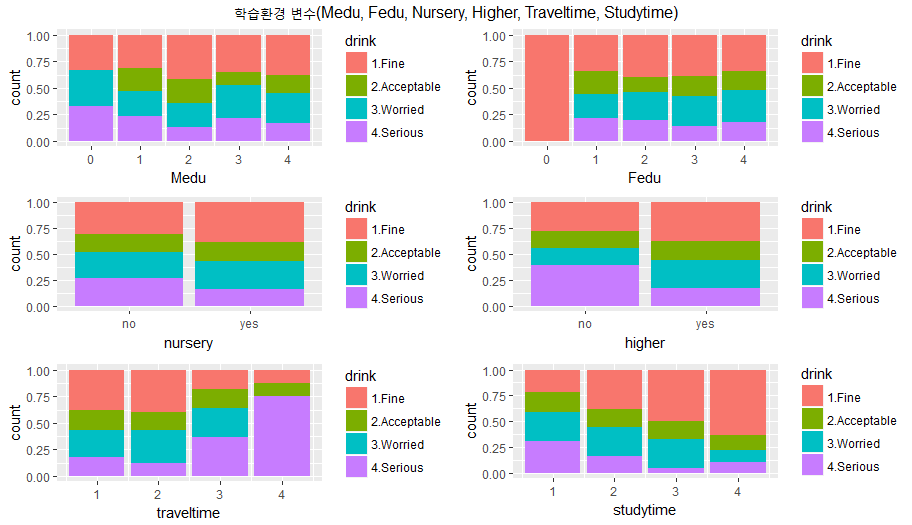
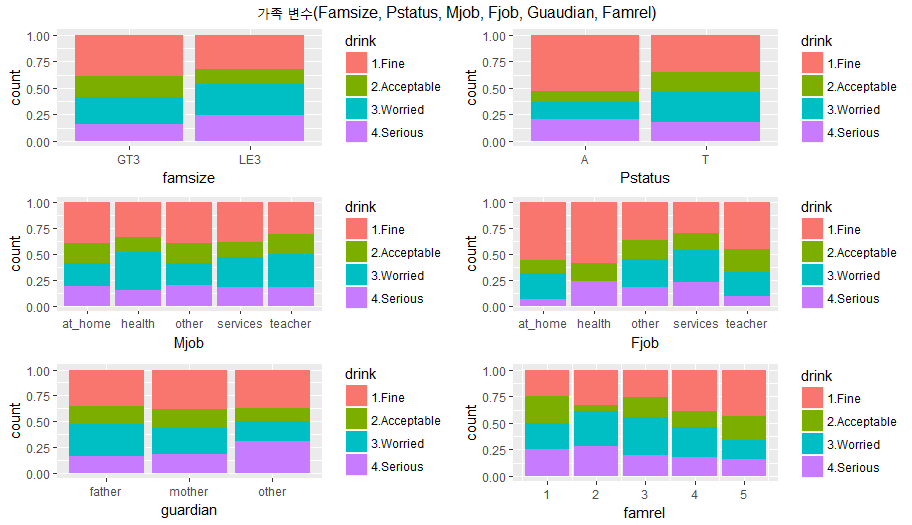
본 데이터는 원래 학생들의 최종 성적을 예측하기 위한 목적으로 공개되었지만, 우리의 분석 목적은 학생들의 음주 수준을 예측하는 것이므로 그 목적에 맞게 데이터셋을 편집하고 몇 가지 파생변수를 추가하였다.

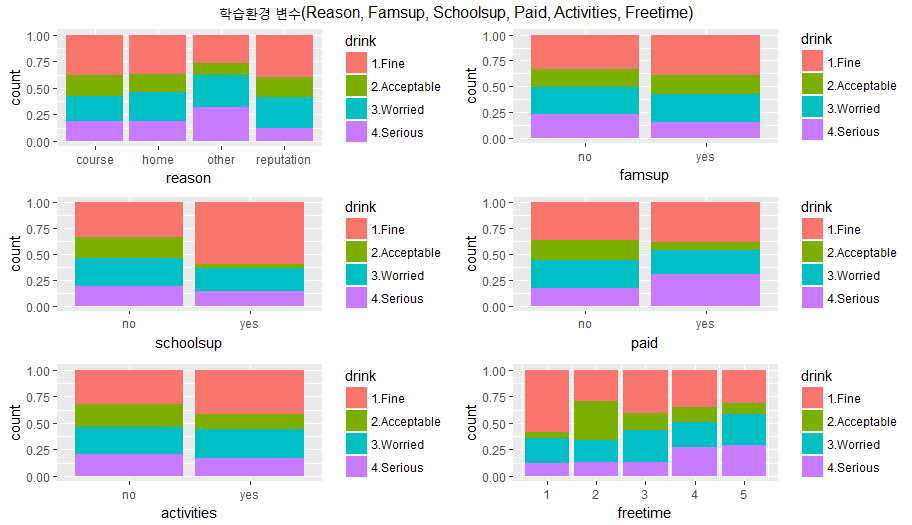
* student-mat.csv과 student-por.csv 두 가지 데이터셋에 모두 포함되어 있는 학생들(382명)을 선별하여 하나의 데이터셋으로 새롭게 구성하였다.
* Grade – portugese와 math의 성적을 각각 표준화한 뒤 평균을 계산하여 학생들의 학업성취도를 파악해보고자 했다.
* Alco – Dalc변수와 Walc변수를 합함으로써 청소년의 음주수준을 알아보려고 한다.
* Drink – Alco 값의 범위를 4등분하여 나타낸 알코올 섭취 수준으로 종속변수로 사용(numeric: from 1 – Fine, 2 – Acceptable, 3 – Worried, 4 – Serious)

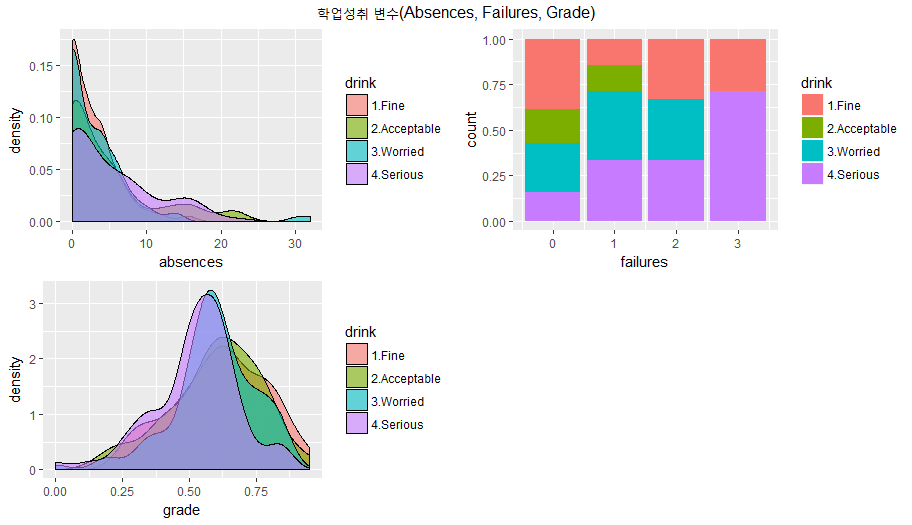
**어떤 변수가 청소년의 알코올 섭취량에 영향을 미칠까?**

독립변수의 class에 따라 종속변수인 drink에 차이가 나타난다면, 해당 독립변수는 drink에 영향을 미치는 변수일 수 있다. 따라서 그러한 차이가 있는지 알아보기 위해 총 29개의 독립변수를 drink 수준에 따라 시각화하여 데이터를 탐색해보았다. (R의 ggplot2패키지를 이용하여 범주형 변수는 100% 누적 막대 그래프, 연속형 변수는 확률밀도 그래프를 그렸다.)









internet, romantic, guardian, farmsup, activities, nursery를 제외한 23개의 변수들은 그래프 상에서 각 변수의 범주별로 음주량에 차이가 나타나는 것으로 보인다.

1) 기본정보 변수 : age, sex, school, health, goout, address (6)

2) 가족 변수 : farmsize, Pstatus, Mjob, Fjob, famrel (5)

3) 학습환경변수 : Medu, Fedu, higher, reason, traveltime, studytime, schoolsup, paid, freetime (9)

4) 학업성취 변수 : absences, failures, grade (3)

이 변수들에서 시각적으로 각 범주별로 음주 수준에 차이를 확인할 수 있었다. 하지만 실제 이러한 차이가 유의미한 수준인지는 통계적인 방법으로 확인해볼 필요가 있다.

**통계적 유의성 검정**

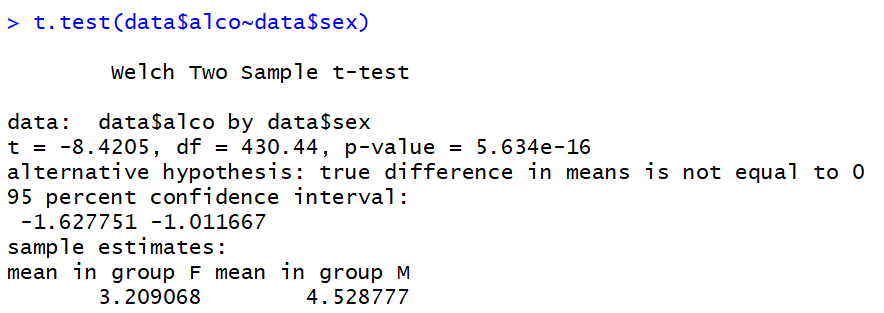
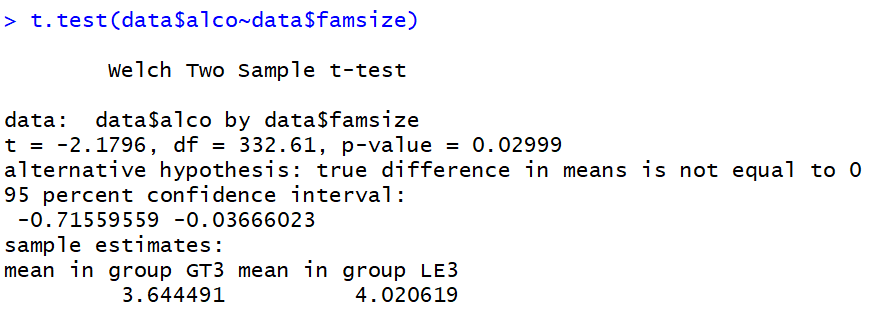
이 과정에서는 범주형 변수인 drink변수로 연속형 변수인 alco 변수를 활용하였다.

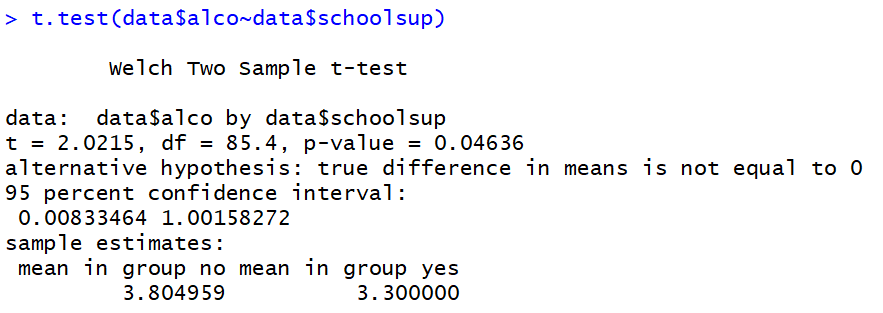
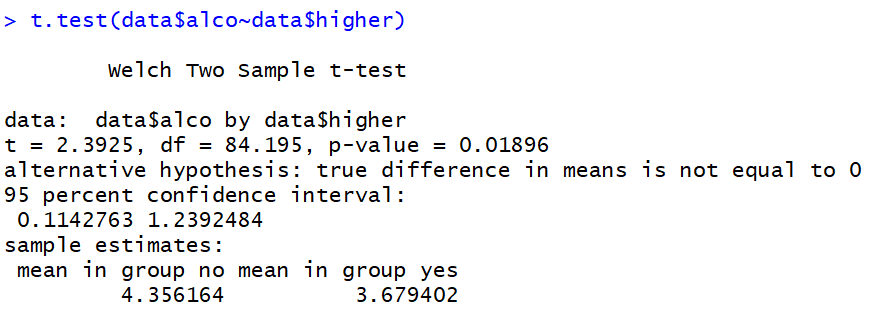
1. 2개의 범주를 가진 변수들의 클래스별 알코올 섭취 수준 차이를 확인하기 위한 t-test

범주형 변수 : 기본 정보 변수(sex, school, address, internet), 가족 변수(famsize, Pstatus), 학습환경변수(higher, schoolsup, paid, activities)

연속형 변수 : alco

2개의 범주로 나뉘어진 범주형 변수들의 경우 t-test를 통해 각 집단에 따라 알코올 섭취 수준의 평균에 차이가 있는지 확인해보았다. 유의수준을 0.05로 설정하였고 총 4개의 변수***(sex, famsize, schoolsup, higher)***들의 p값이 유의수준보다 작게 나타났다. 때문에 이 4개의 변수들을 알코올 섭취 수준에 영향을 미치는 변수로 판단하였다.

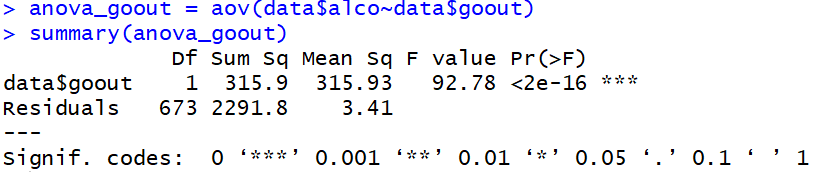
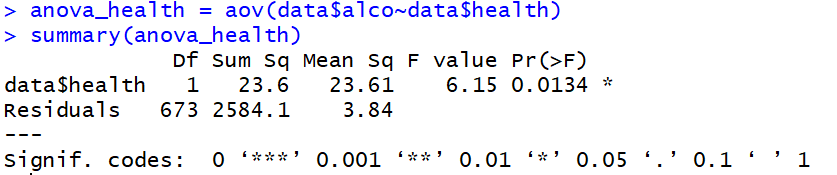
 

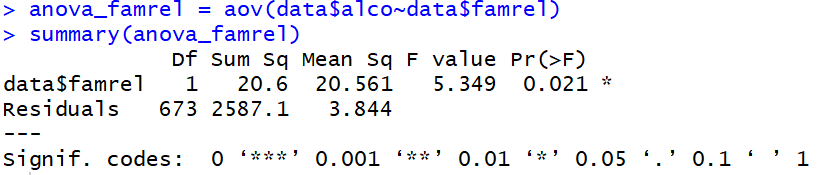
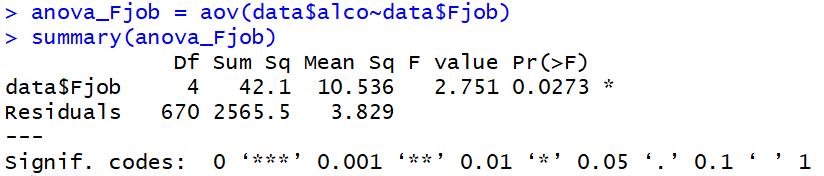
1. 3개 이상의 범주를 가진 변수들의 클래스별 알코올 섭취 수준 차이를 확인하기 위한 anova 분석

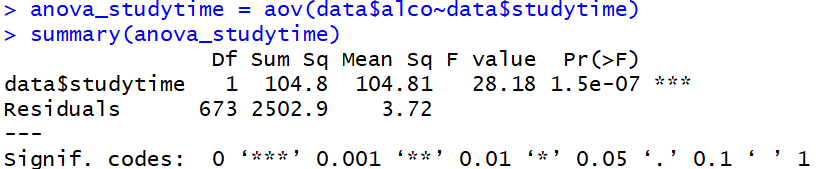
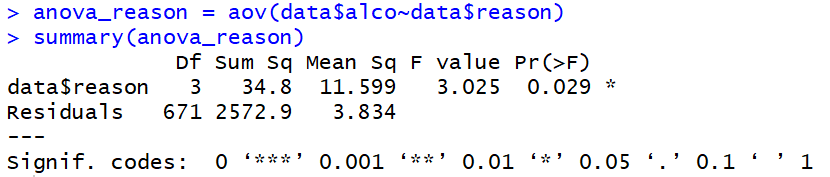
- 범주형 변수 : 기본정보변수(health, gout), 가족/가족관계변수(Mjob, Fjob, famrel) 학습환경변수(Medu, Fedu, reason, traveltime, studytime, freetime) 학습결과변수(failures)

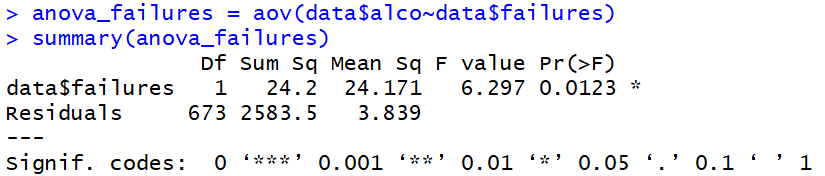
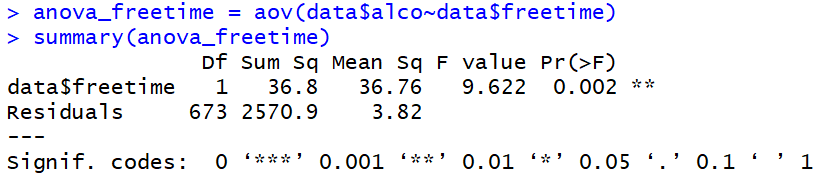
- 연속형 변수: alco

3개 이상의 범주를 가진 변수들의 경우, ANOVA를 통해 각각의 클래스에 따라 알코올 섭취 수준의 평균에 차이가 있는지 확인해보았다. 앞선 t-test와 마찬가지로 유의수준을 5%로 설정했고, 그 결과 총 8개의 변수(***health, gout, Fjob, famrel, reason, studytime, freetime, failures***)들의 p-value가 유의수준보다 작게 나타나 알코올 섭취 수준에 영향을 미치는 변수라고 파악하였다.



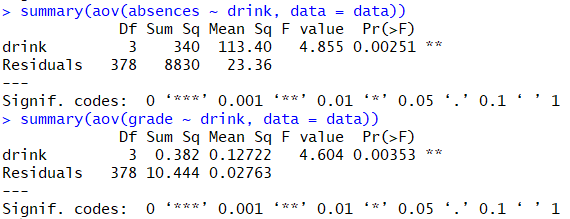






1. ANOVA 알코올 섭취 수준에 따른 연속형 변수들의 평균의 차이를 확인

연속형 변수들이 음주수준에 미치는 영향을 파악하기 위해 이전방법과는 반대로 drink변수의 클래에 따른 absences변수와 grade 변수의 평균 차이를 ANOVA를 통해 파악해 보았다.



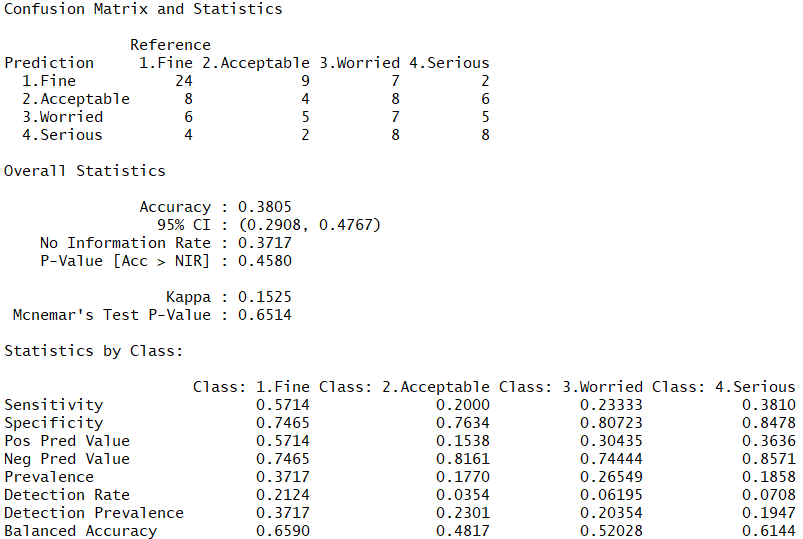
Anova를 실시한 결과, 알코올 섭취 수준에 따라 absences, grade 변수에 차이가 있음을 알 수 있다.

이러한 통계적 검정 결과를 바탕으로 변수 ***sex, famsize, schoolsup, higher, health, goout, Fjob, famrel, reason, studytime, freetime, failures, absences, grade*** 를 청소년의 알코올 섭취 수준에 영향을 미치는 주요 변수로 추출하였다.

**청소년 음주수준 예측 모형**

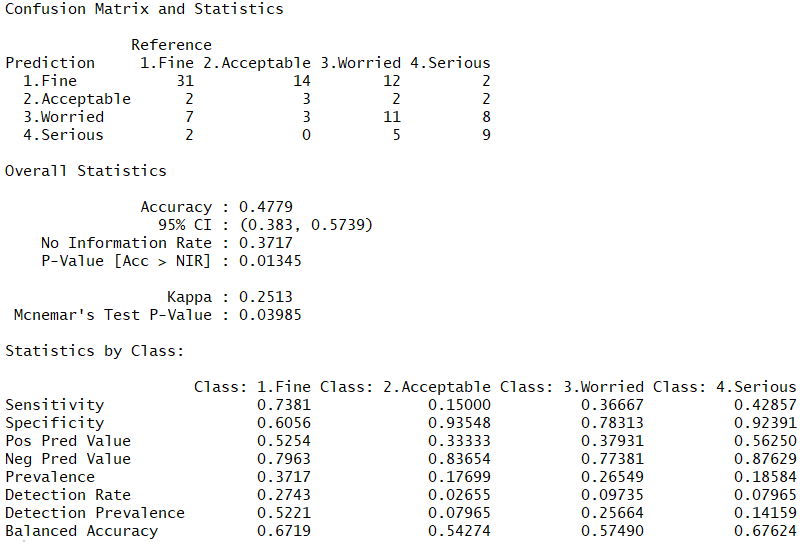
이번 분석의 목적은 청소년의 음주수준을 예측하는 것인 만큼 예측의 성능뿐만이 아니라 그에 대한 설명력 또한 매우 중요하다. 때문에 모형에 대한 설명이 쉬운 의사결정트리 모형을 위주로 분류를 진행해보고자 한다.

의사결정 트리모형

모형을 적용하기에 앞서 데이터셋을 7:3비율로 train, test셋으로 나누어 진행했다. 청소년의 알코올 섭취 수준을 의사결정트리 방법을 적용해 보았다. R의 C5.0함수를 적용하고 예측한 결과는 다음과 같다.

정롹도는 38%, Kappa통계량은 0.15로 예측력이 매우 떨어지는 것으로 나타났다.

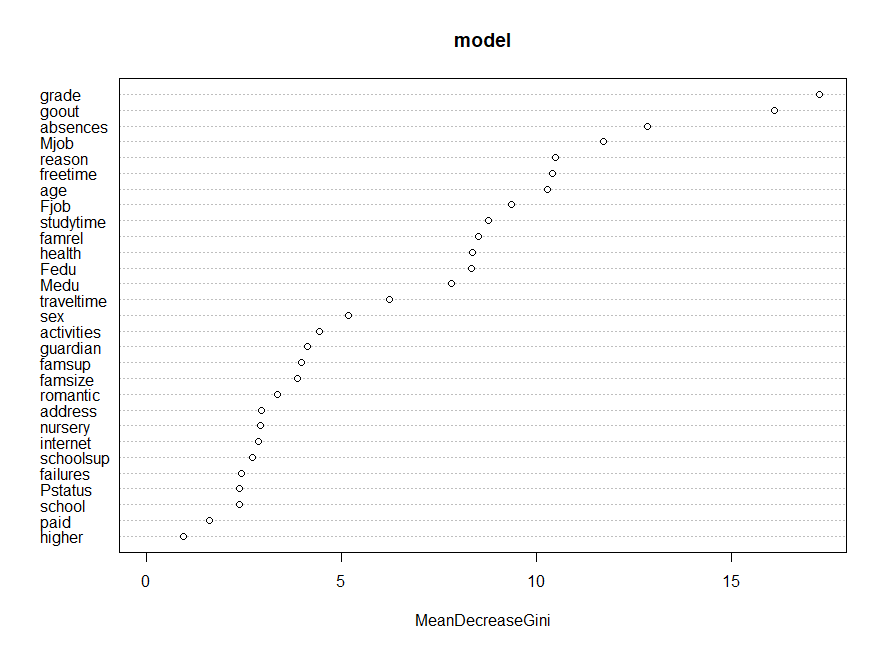
랜덤포레스트

의사결정 트리모형의 예측력을 보완하기 위한 앙상블 방법 중 하나인 랜덤포레스트를 적용해 보았다. R의 RandomForest함수를 사용한 결과는 다음과 같이 나타난다.

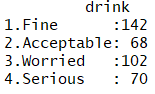
전보다는 조금 예측력이 향상되었지만 아직 부족하다고 생각된다.

**변수선택 및 조정**

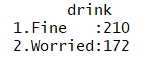
적절한 변수선택이 되지 않아 예측력이 좋지 못한 것이 아닐까 의심이 된다. 그래서 30개의 모든 변수들을 사용한 랜덤포레스트를 통해 변수들의 중요도를 알아보았다,



*Grade, goout, absences,* *Mjob, reason, freetime,* *age, Fjob, studytime, famrel, health,* *Fedu, Medu,* *traveltime, sex*순으로 청소년 음주수준에 영향을 미치는 것으로 나타났다. 그래서 기존에 사용된 변수들에서 ***Fedu, Medu,*** ***Mjob, age, traveltime***를 추가하고 중요도나 낮게 나타난 ***schoolsup, higher, failures*** 변수를 제외하였다.



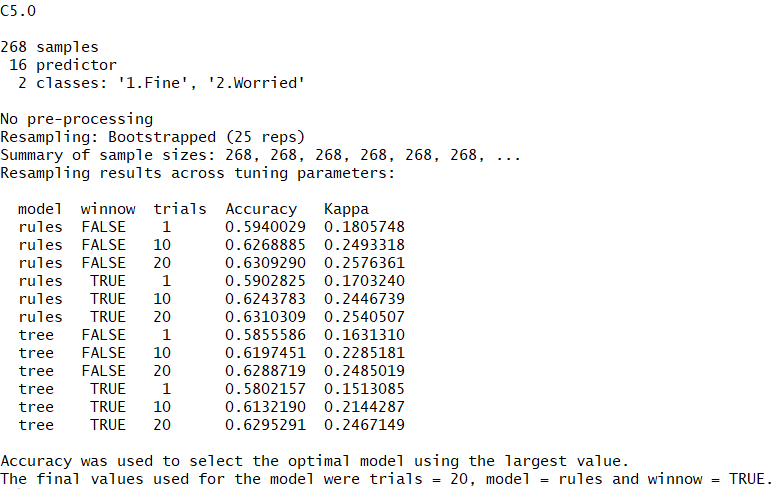
또한 현재 drink변수의 분포는 다음과 같은데 이렇게 범주별 분포가 균일하게 나타나지 않은 것이 모형의 예측력이 낮은 이유로 보인다.

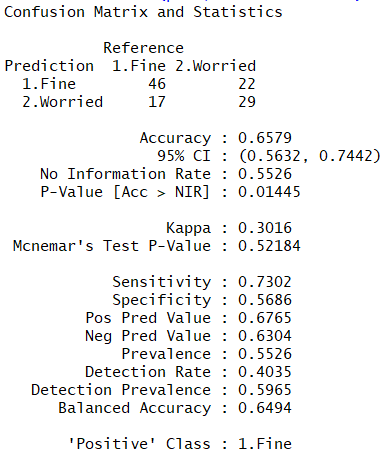
그래서 drink변수의 1,2번 범주와 3,4번 범주를 묶어 2개의 변수로 다시 범주화하니 보다 불균형의 정도가 완화된 것을 확인할 수 있었다.

**개선된 청소년 음주수준 예측 모형**

의사결정트리

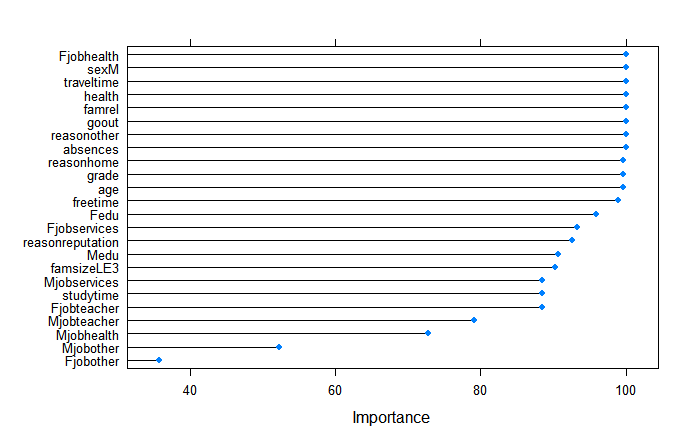
앞선 분석과 마찬가지로 R의 C5.0함수를 통해 의사결정트리모델을 적용한 결과는 다음과 같다. 이번에는 R에서 caret 패키지의 train 함수를 이용해 10fold 교차검증을 적용한 모형을 만들어 보다 개선된 성능을 보일 수 있도록 했다. 아래의 그림은 train 데이터의 정확도를 보여주고 있다.



이 데이터셋의 경우 종속변수 클래스의 불균형이 있기 때문에 위에서 보여주는 것과는 달리 카파통계량이 가장 높은 파라미터가 trias = 20, model = rules, winnow = False인 모형을 적용하였다.

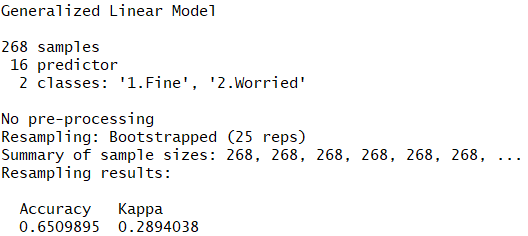
그리고 이를 통해 예측한 결과 정확도는 66%, Kappa 통계량은 0.3으로 이전에 비해 상당히 개선된 것을 확인할 수 있었다.

이 모형에서 주요하게 사용된 변수는 다음과 같이 나타나는데 Fjob, sex, traveltime, health, famrel, goout, reason, absences, grade, age, freetime 변수의 중요도가 매우 높은 것을 확인할 수 있다.

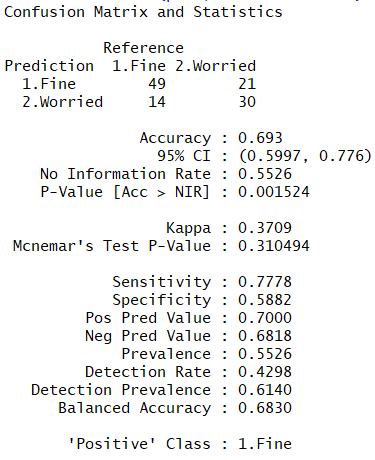


로지스틱 회귀분석

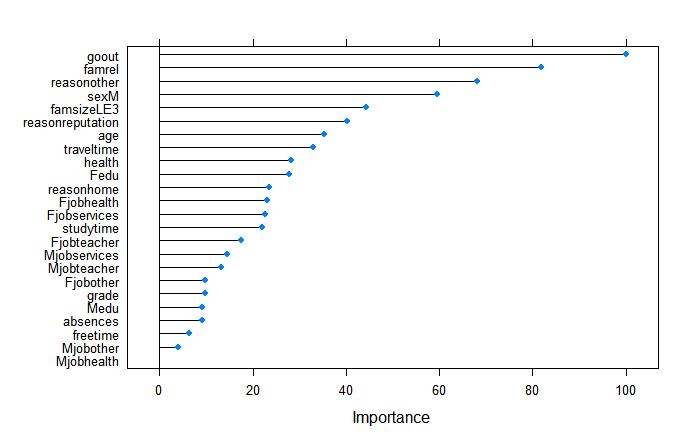
지금까지 트리기반 모형들을 통해 예측을 해보았다. 그렇지만 선택한 16개의 독립변수들과 종속변수를 선형관계에서 설명할 수 있을지 확인하기 위해 로지스틱 회귀분석을 적용해 보았다.



Train데이터의 예측결과만 보더라고 의사결정트리에 비해 조금 더 향상된 결과가 나온 것으로 보인다. 최종예측 결과는 예측 정확도는 69%로 의사결정트리로 예측한 결과에 비해 3%정도 향상되었다. 그리고 kappa통계량은 약 0.07정도 향상된 것으로 나와 로지스틱 회귀모형이 더 적합한 모델이라고 판단된다.



의사결정트리 모형과는 달리 goout변수 하나의 중요도가 다른 변수들에 비해 큰 것으로 나타났다.



**결론**

우리가 만들어낸 최종모형들은 예측 정확도가 66~69%정도인데 이 모형이 사람들의 행동을 예측하는 사회과학적 연구라는 점을 고려하면 상당히 높은 예측률을 보이고 있다고 판단된다. 그래서 과연 이 모형이 우리가 처음 생각했던 학업과 관련된 요인으로 청소년 음주를 설명할 수 있는지 해석해보았다.

우리는 최종모형인 로지스틱 회귀분석을 통해 친구와 함께 나가는 횟수, 가족 친밀도, 학교를 다니는 이유, 성별, 가족규모, 나이, 여행시간, 건강상태, 아버지의 교육수준이 청소년의 음주수준에 크게 영향을 주는 요소라는 것을 알 수 있었다. 이를 보면 청소년 음주는 교우관계, 가족관계에 크게 영향을 받는다는 기존연구와 크게 달라보이지 않는다. 우리의 목표였던 학업과 관련된 스트레스 및 학업 성취도가 청소년기에 음주를 제대로 설명하지 않는 것처럼 보인다.

하지만 예측정확도가 조금 떨어지는 의사결정트리 모형에서는 우리가 원하던 결과를 제공하고 있다. 이 모형에서는 결석의 빈도, 성적, 공부시간, 부모님의 직업이 교사인 변수들의 중요도가 높게 나타나고 있다. 이를 바탕으로 우리는 청소년 음주를 학업과 관련된 요인에 의해서도 예측이 가능하다는 결론을 도출할 수 있었다.

1. http://news.khan.co.kr/kh\_news/khan\_art\_view.html?artid=201711081432002#csidxbf03ebf2e6551d18397786df9147c97 [↑](#footnote-ref-1)