AIDS\_Classification

Shim Soon Yong

2024-05-20

# 데이터셋 설명

캐글의 [AIDS Virus Infection Prediction](https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction)에서 다운로드 받은 데이터셋이다.

## 맥락:

이 데이터셋은 AIDS 진단을 받은 환자에 대한 의료 통계 및 범주형 정보를 포함하고 있다. 1996년에 처음 발표되었다.

### 속성 정보:

* **time**: 실패 또는 중도절단 시간
* **trt**: 치료 지표 (0 = ZDV만, 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI만)
* **age**: 기준 시점의 나이 (연령)
* **wtkg**: 기준 시점의 체중 (킬로그램)
* **hemo**: 혈우병 여부 (0=없음, 1=있음)
* **homo**: 동성애 활동 여부 (0=없음, 1=있음)
* **drugs**: IV 약물 사용 이력 여부 (0=없음, 1=있음)
* **karnof**: 카노프스키 점수 (0-100 점 척도)
* **oprior**: 175일 전 비-ZDV 항레트로바이러스 치료 여부 (0=없음, 1=있음)
* **z30**: 175일 전 30일 동안의 ZDV 사용 여부 (0=없음, 1=있음)
* **preanti**: 175일 전 항레트로바이러스 치료 일수
* **race**: 인종 (0=백인, 1=비백인)
* **gender**: 성별 (0=여성, 1=남성)
* **str2**: 항레트로바이러스 치료 이력 (0=신규, 1=경험 있음)
* **strat**: 항레트로바이러스 치료 역사 분류 (1=‘항레트로바이러스 신규’, 2=‘1주 이상 52주 이하의 항레트로바이러스 치료 경험’, 3=‘52주 이상 치료 경험’)
* **symptom**: 증상 지표 (0=무증상, 1=유증상)
* **treat**: 치료 지표 (0=ZDV만, 1=기타)
* **offtrt**: 96+/-5주 이전에 치료 중단 여부 (0=아니요, 1=예)
* **cd40**: 기준 시점의 CD4 수치
* **cd420**: 20+/-5주 시점의 CD4 수치
* **cd80**: 기준 시점의 CD8 수치
* **cd820**: 20+/-5주 시점의 CD8 수치
* **infected**: AIDS 감염 여부 (0=아니요, 1=예)

#### 추가 변수 정보:

* **개인 정보**: 나이, 체중, 인종, 성별, 성적 활동
* **의료 이력**: 혈우병 여부, IV 약물 사용 이력
* **치료 이력**: ZDV/비-ZDV 치료 이력
* **실험실 결과**: CD4/CD8 수치

# 다변량 분석에 필수적인 패키지 탑재하기

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.2.0 ──  
## ✔ broom 1.0.6 ✔ rsample 1.2.1   
## ✔ dials 1.2.1 ✔ tune 1.2.1   
## ✔ infer 1.0.7 ✔ workflows 1.1.4   
## ✔ modeldata 1.3.0 ✔ workflowsets 1.1.0   
## ✔ parsnip 1.2.1 ✔ yardstick 1.3.1   
## ✔ recipes 1.0.10   
## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Dig deeper into tidy modeling with R at https://www.tmwr.org

library(skimr)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(psych)

##   
## 다음의 패키지를 부착합니다: 'psych'  
##   
## The following objects are masked from 'package:scales':  
##   
## alpha, rescale  
##   
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(MASS)

##   
## 다음의 패키지를 부착합니다: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

tidymodels\_prefer()  
read\_csv("C:/Users/user/OneDrive/바탕 화면/leaRning/Kaggle/data/AIDS\_Classification.csv") -> AIDS

## Rows: 2139 Columns: 23  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (23): time, trt, age, wtkg, hemo, homo, drugs, karnof, oprior, z30, prea...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# 간단하게 훑기(skim) & 데이터 준비

AIDS |> skim()

Data summary

|  |  |
| --- | --- |
| Name | AIDS |
| Number of rows | 2139 |
| Number of columns | 23 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 23 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| time | 0 | 1 | 879.10 | 292.27 | 14 | 727.00 | 997.00 | 1091.00 | 1231.00 | ▁▂▂▃▇ |
| trt | 0 | 1 | 1.52 | 1.13 | 0 | 1.00 | 2.00 | 3.00 | 3.00 | ▇▇▁▇▇ |
| age | 0 | 1 | 35.25 | 8.71 | 12 | 29.00 | 34.00 | 40.00 | 70.00 | ▁▇▆▂▁ |
| wtkg | 0 | 1 | 75.13 | 13.26 | 31 | 66.68 | 74.39 | 82.56 | 159.94 | ▁▇▂▁▁ |
| hemo | 0 | 1 | 0.08 | 0.28 | 0 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| homo | 0 | 1 | 0.66 | 0.47 | 0 | 0.00 | 1.00 | 1.00 | 1.00 | ▅▁▁▁▇ |
| drugs | 0 | 1 | 0.13 | 0.34 | 0 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| karnof | 0 | 1 | 95.45 | 5.90 | 70 | 90.00 | 100.00 | 100.00 | 100.00 | ▁▁▁▅▇ |
| oprior | 0 | 1 | 0.02 | 0.15 | 0 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| z30 | 0 | 1 | 0.55 | 0.50 | 0 | 0.00 | 1.00 | 1.00 | 1.00 | ▆▁▁▁▇ |
| preanti | 0 | 1 | 379.18 | 468.66 | 0 | 0.00 | 142.00 | 739.50 | 2851.00 | ▇▃▁▁▁ |
| race | 0 | 1 | 0.29 | 0.45 | 0 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▃ |
| gender | 0 | 1 | 0.83 | 0.38 | 0 | 1.00 | 1.00 | 1.00 | 1.00 | ▂▁▁▁▇ |
| str2 | 0 | 1 | 0.59 | 0.49 | 0 | 0.00 | 1.00 | 1.00 | 1.00 | ▆▁▁▁▇ |
| strat | 0 | 1 | 1.98 | 0.90 | 1 | 1.00 | 2.00 | 3.00 | 3.00 | ▇▁▃▁▇ |
| symptom | 0 | 1 | 0.17 | 0.38 | 0 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| treat | 0 | 1 | 0.75 | 0.43 | 0 | 1.00 | 1.00 | 1.00 | 1.00 | ▂▁▁▁▇ |
| offtrt | 0 | 1 | 0.36 | 0.48 | 0 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▅ |
| cd40 | 0 | 1 | 350.50 | 118.57 | 0 | 263.50 | 340.00 | 423.00 | 1199.00 | ▂▇▂▁▁ |
| cd420 | 0 | 1 | 371.31 | 144.63 | 49 | 269.00 | 353.00 | 460.00 | 1119.00 | ▃▇▃▁▁ |
| cd80 | 0 | 1 | 986.63 | 480.20 | 40 | 654.00 | 893.00 | 1207.00 | 5011.00 | ▇▅▁▁▁ |
| cd820 | 0 | 1 | 935.37 | 444.98 | 124 | 631.50 | 865.00 | 1146.50 | 6035.00 | ▇▁▁▁▁ |
| infected | 0 | 1 | 0.24 | 0.43 | 0 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |

AIDS %>% mutate(  
 across(c(trt,hemo,homo,drugs,oprior,z30,race,gender,str2,strat,symptom,treat,offtrt,infected),   
 as.factor)) -> AIDS  
AIDS %>% skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 2139 |
| Number of columns | 23 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 14 |
| numeric | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| trt | 0 | 1 | FALSE | 4 | 3: 561, 0: 532, 2: 524, 1: 522 |
| hemo | 0 | 1 | FALSE | 2 | 0: 1959, 1: 180 |
| homo | 0 | 1 | FALSE | 2 | 1: 1414, 0: 725 |
| drugs | 0 | 1 | FALSE | 2 | 0: 1858, 1: 281 |
| oprior | 0 | 1 | FALSE | 2 | 0: 2092, 1: 47 |
| z30 | 0 | 1 | FALSE | 2 | 1: 1177, 0: 962 |
| race | 0 | 1 | FALSE | 2 | 0: 1522, 1: 617 |
| gender | 0 | 1 | FALSE | 2 | 1: 1771, 0: 368 |
| str2 | 0 | 1 | FALSE | 2 | 1: 1253, 0: 886 |
| strat | 0 | 1 | FALSE | 3 | 1: 886, 3: 843, 2: 410 |
| symptom | 0 | 1 | FALSE | 2 | 0: 1769, 1: 370 |
| treat | 0 | 1 | FALSE | 2 | 1: 1607, 0: 532 |
| offtrt | 0 | 1 | FALSE | 2 | 0: 1363, 1: 776 |
| infected | 0 | 1 | FALSE | 2 | 0: 1618, 1: 521 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| time | 0 | 1 | 879.10 | 292.27 | 14 | 727.00 | 997.00 | 1091.00 | 1231.00 | ▁▂▂▃▇ |
| age | 0 | 1 | 35.25 | 8.71 | 12 | 29.00 | 34.00 | 40.00 | 70.00 | ▁▇▆▂▁ |
| wtkg | 0 | 1 | 75.13 | 13.26 | 31 | 66.68 | 74.39 | 82.56 | 159.94 | ▁▇▂▁▁ |
| karnof | 0 | 1 | 95.45 | 5.90 | 70 | 90.00 | 100.00 | 100.00 | 100.00 | ▁▁▁▅▇ |
| preanti | 0 | 1 | 379.18 | 468.66 | 0 | 0.00 | 142.00 | 739.50 | 2851.00 | ▇▃▁▁▁ |
| cd40 | 0 | 1 | 350.50 | 118.57 | 0 | 263.50 | 340.00 | 423.00 | 1199.00 | ▂▇▂▁▁ |
| cd420 | 0 | 1 | 371.31 | 144.63 | 49 | 269.00 | 353.00 | 460.00 | 1119.00 | ▃▇▃▁▁ |
| cd80 | 0 | 1 | 986.63 | 480.20 | 40 | 654.00 | 893.00 | 1207.00 | 5011.00 | ▇▅▁▁▁ |
| cd820 | 0 | 1 | 935.37 | 444.98 | 124 | 631.50 | 865.00 | 1146.50 | 6035.00 | ▇▁▁▁▁ |

각 변수별 분포의 형태가 복잡하다. 복잡하다는 것은 곧 미지의 복잡한 패턴을 따를 수도 있다는 것이다. 패턴 파악을 위해 기본적인 다변량 가설검정들을 실시하겠다.

# 다변량 분석: 가설검정

## 에이즈 감염자 그룹 간 평균 차이 검정하기

# 필요 패키지  
library(Hotelling)

## 필요한 패키지를 로딩중입니다: corpcor

library(biotools)

## ---  
## biotools version 4.2

# 가설검정 모델에 적합한 전처리  
AIDS |> select(-c(trt,hemo,homo,drugs,oprior,z30,race,gender,str2,strat,symptom,treat,offtrt,infected)) -> num.vars  
  
AIDS |> filter(infected == 0) |> select(-c(trt,hemo,homo,drugs,oprior,z30,race,gender,str2,strat,symptom,treat,offtrt,infected)) -> Not.infect  
AIDS |> filter(infected == 1) |> select(-c(trt,hemo,homo,drugs,oprior,z30,race,gender,str2,strat,symptom,treat,offtrt,infected)) -> infect  
boxM(num.vars, group = AIDS$infected)

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: num.vars  
## Chi-Sq (approx.) = 190.58, df = 45, p-value < 2.2e-16

# 가설검정  
cat("Hotelling's T squared test for diffrence between Mean Vectors\n")

## Hotelling's T squared test for diffrence between Mean Vectors

hotelling.test(x = Not.infect, y = infect, var.equal = FALSE) -> T.square  
T.square

## Test stat: 1055.1   
## Numerator df: 9   
## Denominator df: 818.316431318571   
## P-value: 0

Box의 M 검정은 공분산 행렬 간의 동질성(Homogeneity)을 검정한다. 귀무가설은 F검정과 동일하게 두 그룹 간에 분산의 차이가 없다는 것인데, 분산이 그냥 분산이 아니라 여러 개의 변수들에 대한 분산과 공분산을 담은 공분산 행렬을 비교한다. 여기서는 유의확률이 0에 가까워 공분산 행렬이 같다고 할 수 없고, 따라서 호텔링의 검정을 하는 함수의 *var.equal* 옵션에 FALSE를 입력하였다. 호텔링의 검정 결과는 예상대로 유의확률 0으로 에이즈 감염자와 비감염자 간 유의미한 평균 벡터의 차이가 있었다.

## 다변량 분산분석

다변량 분산분석은 여러 개의 평균 벡터의 차이를 여러 개의 그룹 간에 한꺼번에 비교하여 확인할 수 있는 가설 검정법이다.

attach(AIDS)  
manova(cbind(wtkg, preanti, age, cd40, cd420, cd80, cd820, time) ~   
 symptom + infected + homo + hemo + gender + offtrt + oprior + race + str2 + strat + z30 + drugs + karnof) |> summary()

## Df Pillai approx F num Df den Df Pr(>F)   
## symptom 1 0.04942 13.77 8 2118 < 2.2e-16 \*\*\*  
## infected 1 0.45233 218.67 8 2118 < 2.2e-16 \*\*\*  
## homo 1 0.06163 17.39 8 2118 < 2.2e-16 \*\*\*  
## hemo 1 0.10575 31.31 8 2118 < 2.2e-16 \*\*\*  
## gender 1 0.05770 16.21 8 2118 < 2.2e-16 \*\*\*  
## offtrt 1 0.28414 105.08 8 2118 < 2.2e-16 \*\*\*  
## oprior 1 0.02763 7.52 8 2118 6.013e-10 \*\*\*  
## race 1 0.05912 16.64 8 2118 < 2.2e-16 \*\*\*  
## str2 1 0.63584 462.26 8 2118 < 2.2e-16 \*\*\*  
## strat 1 0.51605 282.31 8 2118 < 2.2e-16 \*\*\*  
## z30 1 0.00801 2.14 8 2118 0.029615 \*   
## drugs 1 0.00945 2.53 8 2118 0.009815 \*\*   
## karnof 1 0.00969 2.59 8 2118 0.008130 \*\*   
## Residuals 2125   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

detach(AIDS)

거의 모든 범주형 변수가 유의수준 0.1% 하에서조차 연속형 변수에 대해 차이를 보였다. 그러나 z30 변수와 drug 변수는 비교적 유의성이 낮아 유의수준 1% 하에서만 둘 다 기각된다. 이 원인은 변수를 해석해보면 밝혀질 수도 있다.

z30은 “ZDV in the 30 days prior to 175”의 약자로, 이 문구는 특정 날짜를 기준으로 한 30일 동안의 ZDV 사용을 의미한다. 여기서 ZDV는 아마도 지돌부딘(Zidovudine)의 약어일 가능성이 크다. 지돌부딘은 HIV/AIDS 치료에 사용되는 항레트로바이러스 약물이다. AIDS 치료제를 사용하는 사람이 에이즈 환자 전체의 경향과는 달리 비감염자와 그나마 유사한 특성을 보인다는 것은 얼핏 이해가 안되지만, 이미 에이즈에 걸려서 고생을 하고 행동에 변화가 있었을 것이란 걸 고려하면 어느 정도는 납득할 수 있다. 또한 유의확률이 상대적으로 높을 뿐 여전히 일반적으로 적용되는 유의수준 5% 하에선 귀무가설을 기각하고 유의미한 차이가 있다고 볼 수 있다.

drug는 “history of IV drug use”이며 IV 약물 사용(정맥주사 약물 사용)은 약물을 정맥을 통해 직접 주입하는 것을 의미한다. IV 약물 사용의 목적은 다양하며, 의료적 사용과 비의료적 사용(오남용)으로 구분될 수 있다.

의료적 사용 의료적인 목적으로 IV 약물은 다음과 같은 상황에서 사용:

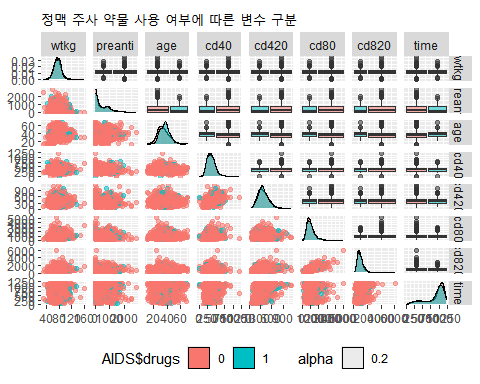
* 신속한 약물 전달: 약물이 빠르게 혈류에 흡수되어 즉각적인 효과를 내야 할 때.
* 고용량 약물 투여: 경구 투여로는 불가능한 고용량의 약물을 투여할 때.
* 입원 치료: 병원에서 환자에게 지속적으로 약물을 공급해야 할 때.
* 수분 및 전해질 보충: 탈수 상태이거나 전해질 불균형이 있을 때.
* 영양 공급: 경구 섭취가 불가능한 환자에게 영양을 공급할 때(TPN, 전 비경구 영양).
* 항생제 투여: 감염 치료를 위해 고용량의 항생제를 투여할 때.
* 진통제 및 진정제 투여: 심한 통증이나 불안, 스트레스를 신속히 완화해야 할 때.
* 항암제 투여: 암 치료를 위해 화학요법 약물을 투여할 때.

비의료적 사용 - 비의료적인 목적으로는 주로 약물 오남용을 의미한다. 이는 불법 약물 또는 처방 약물을 오용하는 것을 포함한다. 주요 사례는 다음과 같다:

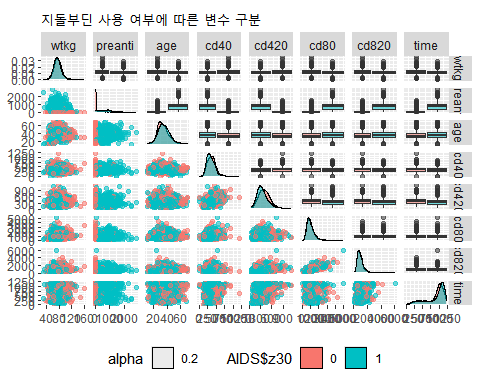
* 불법 약물: 헤로인, 메스암페타민, 코카인 등의 불법 약물을 주사.
* 처방 약물 오용: 처방된 진통제, 진정제 등을 정맥주사로 남용.
* 약물 의존: 주사 약물에 대한 의존성이 생긴 경우.
* 감염 위험: 비위생적인 주사기 사용으로 인한 감염 위험(HIV, B형/C형 간염 등).

에이즈 관련 변수이므로 비의료적 사용에 해당할 것이다. 유의수준 1%에서 평균 벡터의 동일성을 기각할 수 있으므로 적은 차이는 아니지만, 다른 변수들에 비해 상대적으로 덜 유의미하게 차이가 날 이유는 아마 약물 사용에 직접적으로 연관된 지표에 해당하는 변수가 없어서일 것이다.

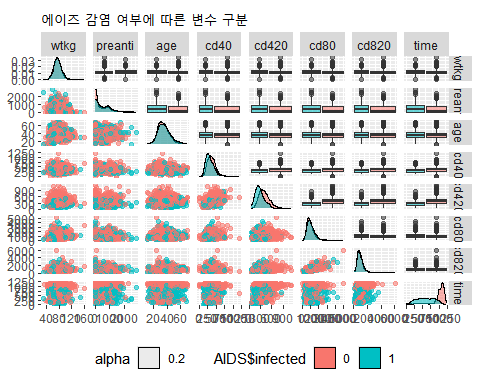
AIDS |> select(wtkg, preanti, age, cd40, cd420, cd80, cd820, time) |>  
 ggpairs(aes(color = AIDS$drugs, alpha = 0.2), title = "정맥 주사 약물 사용 여부에 따른 변수 구분", legend = 1,  
 upper = list(continuous = "box\_no\_facet", combo = "box\_no\_facet", discrete = "count", na = "na"),  
 lower = list(continuous = "points", combo = "facethist", discrete = "facetbar", na ="na"),  
) + theme(legend.position = "bottom")



AIDS |> select(wtkg, preanti, age, cd40, cd420, cd80, cd820, time) |>  
 ggpairs(aes(color = AIDS$z30, alpha = 0.2), title = "지돌부딘 사용 여부에 따른 변수 구분", legend = 1,  
 upper = list(continuous = "box\_no\_facet", combo = "box\_no\_facet", discrete = "count", na = "na"),  
 lower = list(continuous = "points", combo = "facethist", discrete = "facetbar", na = "na")   
)+ theme(legend.position = "bottom")



AIDS |> select(wtkg, preanti, age, cd40, cd420, cd80, cd820, time) |>  
 ggpairs(aes(color = AIDS$infected, alpha = 0.2), title = "에이즈 감염 여부에 따른 변수 구분", legend = 1,  
 upper = list(continuous = "box\_no\_facet", combo = "box\_no\_facet", discrete = "count", na = "na"),  
 lower = list(continuous = "points", combo = "facethist", discrete = "facetbar", na = "na")   
)+ theme(legend.position = "bottom")



산점도 행렬에선 거의 모든 변수가 다 겹쳐서 보일 정도로 유의미한 차이를 발견할 수 없었다. 하지만 대체로 정규분포와 유사한 형태를 띈다는 점은 공통적으로 발견되었다. 에이즈 감염 여부에 따라 구분한 산점도 행렬의 경우에는 *time*변수에서 유의미한 차이를 발견할 수 있었으나, 데이터셋 설명에 따르면 **time: time to failure or censoring**이라고 되어 있다. 즉 실험이 끝나기 얼마 전에 중도절단(censoring)이 많이 발생한 것인데, 관측 시간이 길어질수록 이탈이 자주 발생하는 것은 특별한 사항은 아니다.

# 연속형 데이터 분석

앞선 가설검정에서는 범주형 변수로 분할된 연속형 변수들을 통해 그것들의 차이를 탐구하였다. 이번엔 실제 범주형 변수들을 제거하고 연속형 변수만으로 추론을 진행하여, 연속형 변수들만의 고유한 특성을 탐색해 보겠다.

## 상관분석

상관계수는 0이면 무관, 1이면 양의 상관관계, -1이면 음의 상관관계를 의미한다. 상관계수를 구하는 것은 어렵지 않고 따라서 실질적으로 해야 할 것은 상관계수 검정이다. 상관계수 검정을 통해 변수 간 상관관계가 유의미하게 큰지 확인하면 더 고급 분석을 할 수 있다.

cat("Bartlett's test for sphericity\n")

## Bartlett's test for sphericity

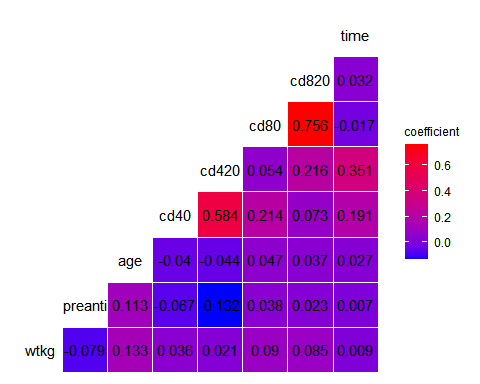
AIDS |> select(wtkg, preanti, age, cd40, cd420, cd80, cd820, time) |>  
 cortest.bartlett()

## R was not square, finding R from data

## $chisq  
## [1] 3764.862  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 28

AIDS |> select(wtkg, preanti, age, cd40, cd420, cd80, cd820, time) |>  
 ggcorr(label = TRUE, label\_round = 3) + scale\_fill\_gradient(low = "blue", high = "red")

## Scale for fill is already present.  
## Adding another scale for fill, which will replace the existing scale.



바틀렛의 구형성 검정에서 유의확률이 매우 낮으므로 구형성 귀무가설을 기각한다. 여기서 구형성이란 상관행렬이 단위행렬과 동일한 성질을 띄는 것을 의미한다. 그리고 상관행렬은 단위행렬과 다를 때 변수 간 상관관계를 나타내기 때문에, 구형성이 기각되었다는 것은 곧 상관관계를 의심할 수 있다는 것이다.

실제 상관행렬을 보아도 숫자가 적지 않아 충분히 연속형 변수 간 상관관계를 의심할 수 있다.

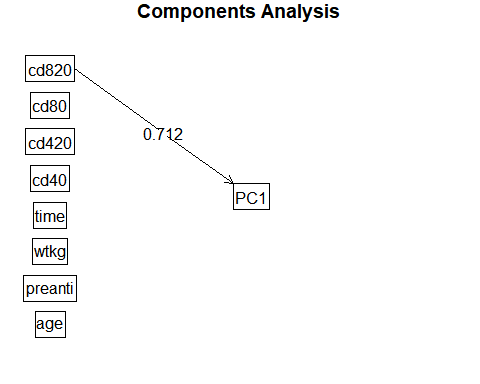
## 주성분 요인분석

주성분 요인분석은 주성분분석 기반으로 요인을 찾아내는 기법으로, 적재량이 급격히 감소하지 않는 개수까지의 주성분 또는 일정 기준선 이상의 주성분을 요인으로 가정하고 분석한다. 이는 다변량 분산분석을 역으로 하는 것으로 이해할 수 있다. 다변량 분산분석은 연속형 변수들을 범주형 변수들로 쪼개서 공분산 행렬을 고려할 때 유의미한 평균 벡터의 차이가 있는지 검정한다. 반대로, 요인분석은 요인이라는 범주형 변수를 연속형 변수가 분산을 통해 쪼개는 구조이다. 그런데 분산은 연속형 변수에 대해 원래 구할 수가 있는 값이다. 따라서 일단 가지고 있는 연속형 데이터의 분산으로 가상의 요인을 구축한 다음 분산분석처럼 변수가 속하는 범주(요인)에 따라 왜 서로 다른 요인에 속하는지 해석하는 것은 충분히 가능하다. 즉, 요인분석을 하면 잠재된 범주형 변수를 구할 수가 있으며, 이를 잠재변수라고 한다.

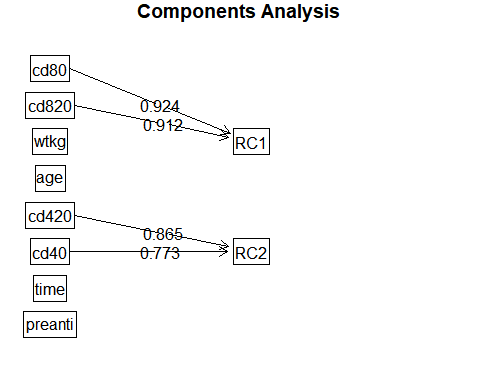
이러한 이유로 주성분 요인분석을 수행하면 실제로 큰 영향을 미칠 법 한 범주형 변수의 한도는 몇 개 까지인지 생각해 볼 수 있다.

for (i in 1:8){  
 AIDS %>% select(wtkg, preanti, age, cd40, cd420, cd80, cd820, time) |>  
 principal(nfactors = i, rotate = "varimax") -> temp  
   
 print(temp)  
  
 fa.diagram(temp, simple = FALSE, cut = 0.7, digit = 3)  
}

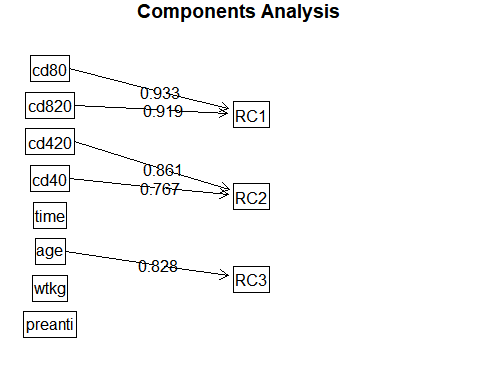
## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PC1 h2 u2 com  
## wtkg 0.17 0.02813 0.97 1  
## preanti -0.10 0.00919 0.99 1  
## age 0.02 0.00051 1.00 1  
## cd40 0.65 0.42425 0.58 1  
## cd420 0.68 0.46729 0.53 1  
## cd80 0.69 0.47736 0.52 1  
## cd820 0.71 0.50647 0.49 1  
## time 0.36 0.13042 0.87 1  
##   
## PC1  
## SS loadings 2.04  
## Proportion Var 0.26  
##   
## Mean item complexity = 1  
## Test of the hypothesis that 1 component is sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.17   
## with the empirical chi square 3274.71 with prob < 0   
##   
## Fit based upon off diagonal values = 0.39



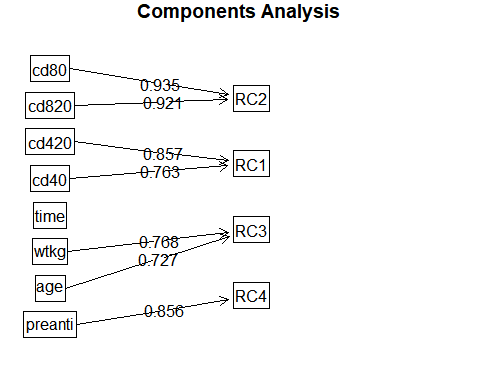
## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC1 RC2 h2 u2 com  
## wtkg 0.22 0.01 0.049 0.95 1.0  
## preanti 0.11 -0.26 0.080 0.92 1.3  
## age 0.17 -0.15 0.053 0.95 2.0  
## cd40 0.18 0.77 0.629 0.37 1.1  
## cd420 0.14 0.87 0.768 0.23 1.1  
## cd80 0.92 0.01 0.855 0.15 1.0  
## cd820 0.91 0.05 0.835 0.17 1.0  
## time -0.04 0.58 0.343 0.66 1.0  
##   
## RC1 RC2  
## SS loadings 1.83 1.78  
## Proportion Var 0.23 0.22  
## Cumulative Var 0.23 0.45  
## Proportion Explained 0.51 0.49  
## Cumulative Proportion 0.51 1.00  
##   
## Mean item complexity = 1.2  
## Test of the hypothesis that 2 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.1   
## with the empirical chi square 1215.89 with prob < 6.7e-252   
##   
## Fit based upon off diagonal values = 0.77



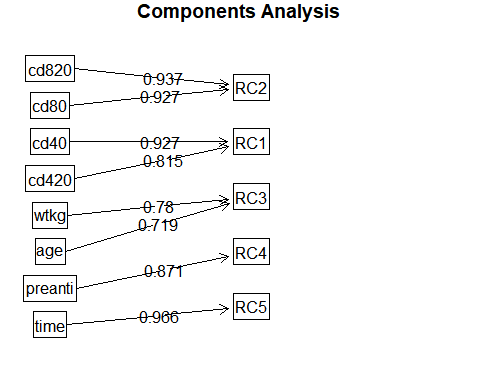
## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC1 RC2 RC3 h2 u2 com  
## wtkg 0.10 0.11 0.50 0.27 0.73 1.2  
## preanti 0.02 -0.18 0.41 0.20 0.80 1.4  
## age -0.03 0.02 0.83 0.69 0.31 1.0  
## cd40 0.17 0.77 -0.12 0.63 0.37 1.1  
## cd420 0.12 0.86 -0.12 0.77 0.23 1.1  
## cd80 0.93 0.03 0.07 0.88 0.12 1.0  
## cd820 0.92 0.08 0.07 0.86 0.14 1.0  
## time -0.12 0.63 0.17 0.45 0.55 1.2  
##   
## RC1 RC2 RC3  
## SS loadings 1.78 1.78 1.16  
## Proportion Var 0.22 0.22 0.15  
## Cumulative Var 0.22 0.45 0.59  
## Proportion Explained 0.38 0.38 0.25  
## Cumulative Proportion 0.38 0.75 1.00  
##   
## Mean item complexity = 1.1  
## Test of the hypothesis that 3 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.12   
## with the empirical chi square 1765.51 with prob < 0   
##   
## Fit based upon off diagonal values = 0.67



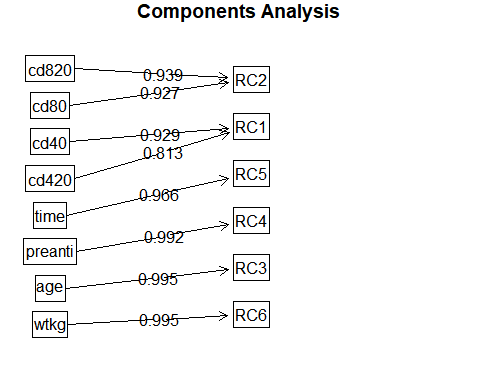
## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC2 RC1 RC3 RC4 h2 u2 com  
## wtkg 0.09 0.00 0.77 -0.37 0.73 0.27 1.5  
## preanti 0.06 -0.06 -0.01 0.86 0.74 0.26 1.0  
## age 0.00 0.01 0.73 0.40 0.69 0.31 1.6  
## cd40 0.16 0.76 -0.05 -0.14 0.63 0.37 1.2  
## cd420 0.11 0.86 -0.04 -0.15 0.77 0.23 1.1  
## cd80 0.94 0.04 0.04 0.02 0.88 0.12 1.0  
## cd820 0.92 0.08 0.04 0.02 0.86 0.14 1.0  
## time -0.12 0.66 0.09 0.21 0.50 0.50 1.3  
##   
## RC2 RC1 RC3 RC4  
## SS loadings 1.78 1.77 1.13 1.11  
## Proportion Var 0.22 0.22 0.14 0.14  
## Cumulative Var 0.22 0.44 0.59 0.73  
## Proportion Explained 0.31 0.30 0.20 0.19  
## Cumulative Proportion 0.31 0.61 0.81 1.00  
##   
## Mean item complexity = 1.2  
## Test of the hypothesis that 4 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.12   
## with the empirical chi square 1672.18 with prob < 0   
##   
## Fit based upon off diagonal values = 0.69



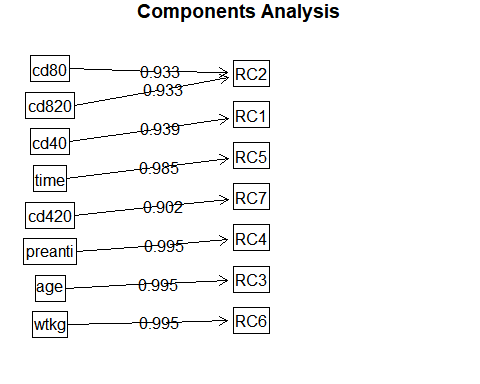
## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC2 RC1 RC3 RC4 RC5 h2 u2 com  
## wtkg 0.10 -0.01 0.78 -0.34 0.01 0.74 0.265 1.4  
## preanti 0.05 -0.08 -0.04 0.87 0.01 0.77 0.230 1.0  
## age -0.02 0.00 0.72 0.45 0.01 0.72 0.280 1.7  
## cd40 0.08 0.93 0.01 0.00 -0.03 0.87 0.135 1.0  
## cd420 0.09 0.81 -0.02 -0.11 0.32 0.78 0.216 1.4  
## cd80 0.93 0.10 0.05 0.04 -0.08 0.88 0.121 1.0  
## cd820 0.94 0.06 0.03 0.00 0.07 0.89 0.114 1.0  
## time -0.01 0.17 0.02 0.02 0.97 0.96 0.038 1.1  
##   
## RC2 RC1 RC3 RC4 RC5  
## SS loadings 1.76 1.57 1.13 1.09 1.05  
## Proportion Var 0.22 0.20 0.14 0.14 0.13  
## Cumulative Var 0.22 0.42 0.56 0.69 0.83  
## Proportion Explained 0.27 0.24 0.17 0.17 0.16  
## Cumulative Proportion 0.27 0.50 0.68 0.84 1.00  
##   
## Mean item complexity = 1.2  
## Test of the hypothesis that 5 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.1   
## with the empirical chi square 1180.84 with prob < NA   
##   
## Fit based upon off diagonal values = 0.78



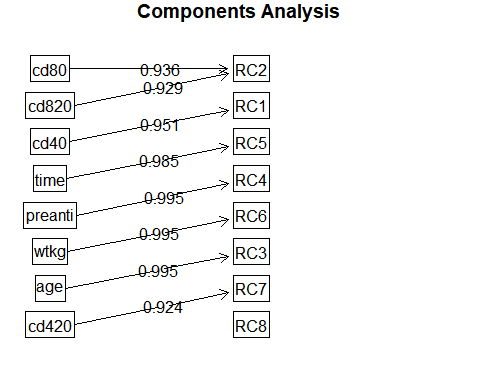
## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC2 RC1 RC5 RC4 RC3 RC6 h2 u2 com  
## wtkg 0.06 0.02 0.01 -0.04 0.07 0.99 1.00 6.6e-04 1.0  
## preanti 0.02 -0.07 0.01 0.99 0.06 -0.04 0.99 6.5e-03 1.0  
## age 0.03 -0.03 0.01 0.06 1.00 0.07 1.00 2.1e-05 1.0  
## cd40 0.07 0.93 -0.03 0.02 -0.02 0.03 0.87 1.3e-01 1.0  
## cd420 0.09 0.81 0.32 -0.12 -0.02 -0.01 0.78 2.2e-01 1.4  
## cd80 0.93 0.10 -0.08 0.04 0.02 0.04 0.88 1.2e-01 1.0  
## cd820 0.94 0.05 0.07 -0.01 0.01 0.03 0.89 1.1e-01 1.0  
## time -0.01 0.17 0.97 0.02 0.02 0.01 0.96 3.8e-02 1.1  
##   
## RC2 RC1 RC5 RC4 RC3 RC6  
## SS loadings 1.76 1.57 1.05 1.01 1.00 1.00  
## Proportion Var 0.22 0.20 0.13 0.13 0.12 0.12  
## Cumulative Var 0.22 0.42 0.55 0.67 0.80 0.92  
## Proportion Explained 0.24 0.21 0.14 0.14 0.14 0.14  
## Cumulative Proportion 0.24 0.45 0.59 0.73 0.86 1.00  
##   
## Mean item complexity = 1.1  
## Test of the hypothesis that 6 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.05   
## with the empirical chi square 318.61 with prob < NA   
##   
## Fit based upon off diagonal values = 0.94



## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC2 RC1 RC5 RC7 RC4 RC3 RC6 h2 u2 com  
## wtkg 0.05 0.02 0.00 0.00 -0.04 0.07 1.00 1.00 5.9e-06 1.0  
## preanti 0.02 -0.03 0.01 -0.06 1.00 0.06 -0.04 1.00 8.0e-05 1.0  
## age 0.03 -0.02 0.01 -0.02 0.06 1.00 0.07 1.00 6.0e-06 1.0  
## cd40 0.08 0.94 0.08 0.29 -0.03 -0.02 0.02 0.98 1.9e-02 1.2  
## cd420 0.08 0.34 0.19 0.90 -0.07 -0.02 0.00 0.98 2.2e-02 1.4  
## cd80 0.93 0.23 -0.01 -0.14 0.01 0.02 0.04 0.94 5.6e-02 1.2  
## cd820 0.93 -0.11 0.00 0.24 0.02 0.01 0.04 0.94 5.6e-02 1.2  
## time 0.00 0.07 0.99 0.15 0.01 0.02 0.00 1.00 2.7e-04 1.1  
##   
## RC2 RC1 RC5 RC7 RC4 RC3 RC6  
## SS loadings 1.76 1.07 1.01 1.00 1.00 1.00 1.00  
## Proportion Var 0.22 0.13 0.13 0.13 0.13 0.13 0.12  
## Cumulative Var 0.22 0.35 0.48 0.61 0.73 0.86 0.98  
## Proportion Explained 0.22 0.14 0.13 0.13 0.13 0.13 0.13  
## Cumulative Proportion 0.22 0.36 0.49 0.62 0.75 0.87 1.00  
##   
## Mean item complexity = 1.1  
## Test of the hypothesis that 7 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.02   
## with the empirical chi square 35.3 with prob < NA   
##   
## Fit based upon off diagonal values = 0.99



## Principal Components Analysis  
## Call: principal(r = AIDS %>% select(wtkg, preanti, age, cd40, cd420,   
## cd80, cd820, time), nfactors = i, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC2 RC1 RC5 RC4 RC6 RC3 RC7 RC8 h2 u2 com  
## wtkg 0.05 0.01 0.00 -0.04 1.00 0.07 0.01 0.00 1 -5.3e-15 1.0  
## preanti 0.02 -0.03 0.01 1.00 -0.04 0.06 -0.06 0.00 1 4.6e-15 1.0  
## age 0.02 -0.02 0.01 0.06 0.07 1.00 -0.02 0.00 1 -1.8e-15 1.0  
## cd40 0.08 0.95 0.08 -0.03 0.02 -0.02 0.28 -0.01 1 0.0e+00 1.2  
## cd420 0.08 0.31 0.19 -0.07 0.01 -0.02 0.92 0.02 1 -2.2e-16 1.3  
## cd80 0.94 0.16 -0.02 0.02 0.04 0.02 -0.06 -0.30 1 -2.2e-16 1.3  
## cd820 0.93 -0.05 0.01 0.01 0.04 0.01 0.17 0.32 1 0.0e+00 1.3  
## time 0.00 0.07 0.98 0.01 0.00 0.02 0.16 0.00 1 1.4e-15 1.1  
##   
## RC2 RC1 RC5 RC4 RC6 RC3 RC7 RC8  
## SS loadings 1.76 1.04 1.01 1.00 1.00 1.00 0.99 0.20  
## Proportion Var 0.22 0.13 0.13 0.13 0.13 0.13 0.12 0.02  
## Cumulative Var 0.22 0.35 0.48 0.60 0.73 0.85 0.98 1.00  
## Proportion Explained 0.22 0.13 0.13 0.13 0.13 0.13 0.12 0.02  
## Cumulative Proportion 0.22 0.35 0.48 0.60 0.73 0.85 0.98 1.00  
##   
## Mean item complexity = 1.2  
## Test of the hypothesis that 8 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0   
## with the empirical chi square 0 with prob < NA   
##   
## Fit based upon off diagonal values = 1



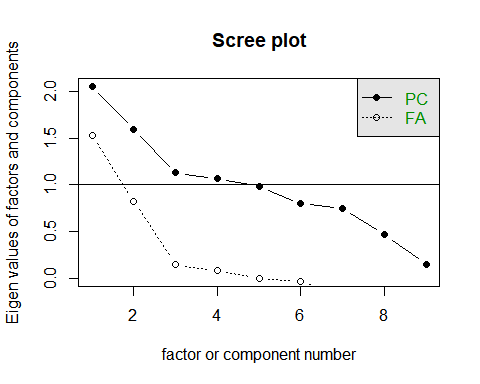
모든 가능한 요인분석을 한 결과, 최적의 요인 수는 요인 적재량에 따라 7개로 확인되었다.

그에 비해 범주형 변수들은 *symptom, infected, homo, hemo, gender, offtrt, oprior, race, str2, strat, z30, drugs, karnof*로 13개나 된다. 요인 적재량이 0.9 이상으로 매우 크며 요인 7개 하에서 괜찮게 분배되었으므로 요인의 분산 적재량 자체의 분산을 최대화하는 VARIMAX 알고리즘이 잘 작동한 것으로도 해석할 수 있다. 이는 곧 범주형 변수들 중에 여기서 구한 요인과 같은 작용을 하는, 원본 데이터의 정보가 손실되지 않으면서도 새로운 역할을 할 변수 집합을 찾아낼 수도 있다는 뜻이다. 결론적으로 원본 데이터를 그대로 사용하는 것을 넘어선 특성 설계가 유의미할 수 있음을 시사한다.

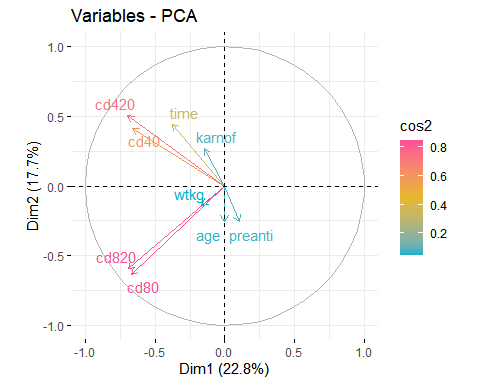
### 주성분 요인분석 시각화

*factoextra* 패키지를 사용하면 쉽게 주성분 분석을 시각화할 수 있다.

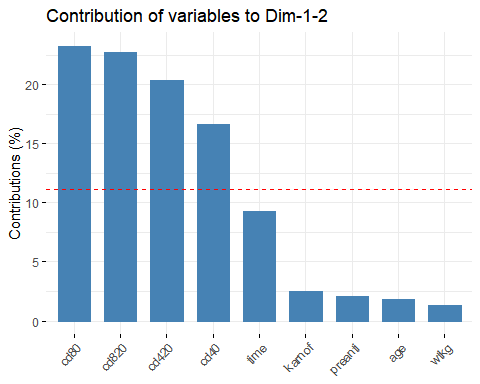
prcomp(num.vars, scale. = TRUE) -> pca.result  
  
scree(num.vars)



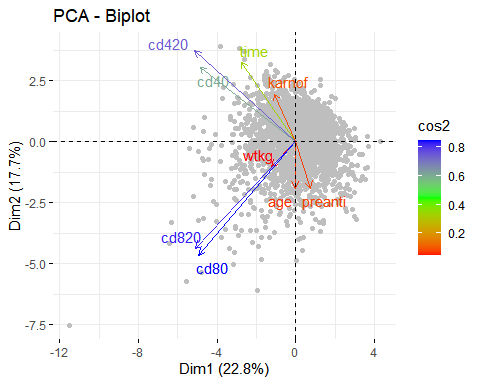
pca.result |>  
 fviz\_pca\_var(axes = c(1, 2), choice = "var", col.var = "cos2", gradient.cols = c("#00AFCC", "#E7B828", "#FD4E97"), repel = TRUE)



pca.result |>  
 fviz\_contrib(axes = c(1, 2), choice = "var")



pca.result |>  
 fviz\_pca\_biplot(col.ind = "grey", geom.ind = "point", choice = "var",   
 col.var = "cos2", gradient.cols = c("red", "green", "blue"), repel = TRUE)



인구통계학적 변수(wtkg, age, preanti)들을 제외한 거의 모든 변수의 분산이 비슷한 방향을 가리키고 있었고, 데이터 포인트가 뚜렷하게 두 군집으로 나뉘어져 있다. 이러한 유형의 데이터에는 선형적인 분류 모형이나 군집 모형이 적절하다.

# 판별분석

## 로지스틱 판별분석

일반화선형모형은 복잡한 데이터 중에서도 정규분포를 따르지 않는 데이터를 다루는 가장 기초적인 도구이다. 결과변수인 감염여부(*infected*)가 0과 1의 값을 가지는 이진변수이므로 그것들 중 로지스틱 회귀를 실시해 적합도를 확인해 볼 수 있다. 로지스틱 회귀는 판별분석에서도 자주 사용되기 때문에 괜찮은 결과를 기대해 볼 수 있다.

glm(infected ~ ., family = binomial(link = "logit"), data = AIDS) -> logit.model   
logit.model |> summary()

##   
## Call:  
## glm(formula = infected ~ ., family = binomial(link = "logit"),   
## data = AIDS)  
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 7.502e+00 1.331e+00 5.637 1.73e-08 \*\*\*  
## time -6.315e-03 3.338e-04 -18.920 < 2e-16 \*\*\*  
## trt1 -1.617e-01 1.990e-01 -0.812 0.41653   
## trt2 -8.874e-02 1.946e-01 -0.456 0.64830   
## trt3 -2.134e-01 1.899e-01 -1.124 0.26115   
## age 1.620e-02 8.114e-03 1.996 0.04593 \*   
## wtkg 2.492e-03 5.479e-03 0.455 0.64928   
## hemo1 -5.741e-01 3.363e-01 -1.707 0.08776 .   
## homo1 -7.057e-03 2.439e-01 -0.029 0.97692   
## drugs1 -4.946e-01 2.307e-01 -2.144 0.03204 \*   
## karnof -2.827e-02 1.187e-02 -2.381 0.01727 \*   
## oprior1 4.247e-01 4.811e-01 0.883 0.37741   
## z301 5.778e-01 3.559e-01 1.624 0.10445   
## preanti 3.908e-04 2.860e-04 1.366 0.17178   
## race1 -5.534e-01 1.722e-01 -3.213 0.00131 \*\*   
## gender1 1.829e-01 2.823e-01 0.648 0.51711   
## str21 -2.646e-01 4.464e-01 -0.593 0.55344   
## strat2 -1.234e-01 2.729e-01 -0.452 0.65115   
## strat3 NA NA NA NA   
## symptom1 4.217e-01 1.742e-01 2.421 0.01546 \*   
## treat1 NA NA NA NA   
## offtrt1 -1.896e+00 1.996e-01 -9.500 < 2e-16 \*\*\*  
## cd40 1.124e-03 8.281e-04 1.357 0.17471   
## cd420 -5.423e-03 8.275e-04 -6.554 5.61e-11 \*\*\*  
## cd80 -5.965e-05 2.417e-04 -0.247 0.80506   
## cd820 5.464e-04 2.661e-04 2.053 0.04006 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2375.0 on 2138 degrees of freedom  
## Residual deviance: 1395.8 on 2115 degrees of freedom  
## AIC: 1443.8  
##   
## Number of Fisher Scoring iterations: 6

logit.model |> predict(newdata = AIDS |> select(-infected), type = "response") -> logit.pred  
  
cat("Predicted infection rate: ")

## Predicted infection rate:

mean(logit.pred >= 0.5)

## [1] 0.1954184

cat("\nTrue infection rate: ")

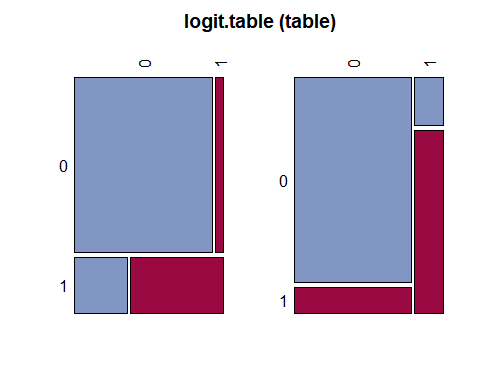
##   
## True infection rate:

mean(AIDS$infected != 0)

## [1] 0.2435718

table(AIDS$infected, as.numeric(logit.pred >= 0.5)) -> logit.table   
  
library(DescTools)  
Desc(logit.table, digits = 2)

## ------------------------------------------------------------------------------   
## logit.table (table)  
##   
## Summary:   
## n: 2'139, rows: 2, columns: 2  
##   
## Pearson's Chi-squared test (cont. adj):  
## X-squared = 844, df = 1, p-value < 2.2e-16  
## Fisher's exact test p-value < 2.2e-16  
## McNemar's chi-squared = 37.56, df = 1, p-value = 8.866e-10  
##   
## estimate lwr.ci upr.ci'  
##   
## odds ratio 30.6570 23.1665 40.5696  
## rel. risk (col1) 2.5947 2.3153 2.9078  
## rel. risk (col2) 0.0846 0.0683 0.1049  
## 0.5815 0.5379 0.6232  
##   
##   
## Contingency Coeff. 0.533  
## Cramer's V 0.630  
## Kendall Tau-b 0.630  
##   
##   
## 0 1 Sum  
##   
## 0 freq 1'531 87 1'618  
## perc 71.58% 4.07% 75.64%  
## p.row 94.62% 5.38% .  
## p.col 88.96% 20.81% .  
##   
## 1 freq 190 331 521  
## perc 8.88% 15.47% 24.36%  
## p.row 36.47% 63.53% .  
## p.col 11.04% 79.19% .  
##   
## Sum freq 1'721 418 2'139  
## perc 80.46% 19.54% 100.00%  
## p.row . . .  
## p.col . . .  
##   
##   
## ----------  
## ' 95% conf. level



로지스틱 회귀모형은 별다른 특성 설계와 전처리 없이도 전체 데이터에서 충분한 예측 정확도를 보였다. 이는 종속변수 *infected*와 독립변수들 사이에 선명한 선형관계가 있음을 증명한다. 그리고 변수들의 유의확률 간에 유의미한 차이가 보이므로, 보다 좋은 특성을 추출하여 계산을 효율화하고 예측 정확도를 올릴 수도 있을 것이다.

## 선형 판별 분석

선형 판별 분석은 마할라노비스 거리 기반 판별과 정준판별함수 기반 판별을 배웠다. 로지스틱 회귀를 통해 충분히 선형 분류가 가능함을 확인하였으므로, 선형 판별 분석도 좋은 성능을 보일 것임을 예상할 수 있다. 여기서는 그중 정준판별함수 기반 판별 분석을 해볼 것이다.

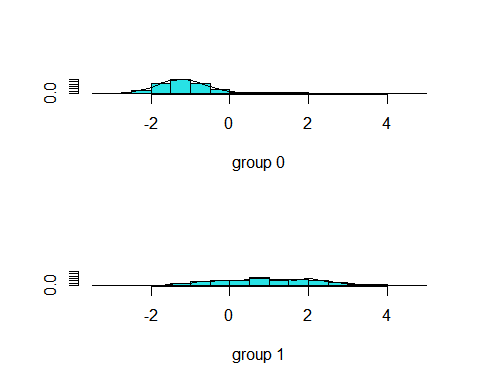
AIDS |>  
 lda(infected ~ . , data = \_) -> lda.result

## Warning in lda.default(x, grouping, ...): 변수들이 공선형관계에 있습니다

lda.result

## Call:  
## lda(infected ~ ., data = AIDS)  
##   
## Prior probabilities of groups:  
## 0 1   
## 0.7564282 0.2435718   
##   
## Group means:  
## time trt1 trt2 trt3 age wtkg hemo1 homo1  
## 0 974.4388 0.2589617 0.2564895 0.2676143 34.90049 75.00338 0.08590853 0.6464771  
## 1 583.0115 0.1976967 0.2092131 0.2456814 36.32821 75.50398 0.07869482 0.7063340  
## drugs1 karnof oprior1 z301 preanti race1 gender1 str21  
## 0 0.1409147 95.7911 0.01854141 0.5148331 345.0229 0.302843 0.8182942 0.5512979  
## 1 0.1017274 94.3762 0.03262956 0.6602687 485.2399 0.243762 0.8579655 0.6928983  
## strat2 strat3 symptom1 treat1 offtrt1 cd40 cd420 cd80  
## 0 0.1897404 0.3615575 0.1452410 0.7830655 0.3374536 362.9895 399.6904 968.8276  
## 1 0.1976967 0.4952015 0.2591171 0.6525912 0.4414587 311.7179 283.1612 1041.9060  
## cd820  
## 0 930.4487  
## 1 950.6526  
##   
## Coefficients of linear discriminants:  
## LD1  
## time -4.345803e-03  
## trt1 -2.502592e-02  
## trt2 -1.382602e-02  
## trt3 -3.627179e-02  
## age 8.138893e-03  
## wtkg 1.714606e-03  
## hemo1 -2.095286e-01  
## homo1 5.828603e-02  
## drugs1 -2.339015e-01  
## karnof -1.305396e-02  
## oprior1 2.346161e-01  
## z301 3.427684e-01  
## preanti 2.231705e-04  
## race1 -2.977894e-01  
## gender1 3.832644e-02  
## str21 -8.292298e-02  
## strat2 -7.326131e-02  
## strat3 -3.594128e-02  
## symptom1 2.270796e-01  
## treat1 -7.710698e-02  
## offtrt1 -9.773230e-01  
## cd40 2.352375e-04  
## cd420 -2.302636e-03  
## cd80 -4.191182e-05  
## cd820 3.115026e-04

plot(lda.result, type = "both")



lda.pred <- predict(lda.result, AIDS)  
  
cat("Predicted infection rate: ")

## Predicted infection rate:

mean(lda.pred$class != 0)

## [1] 0.2071061

cat("\nTrue infection rate: ")

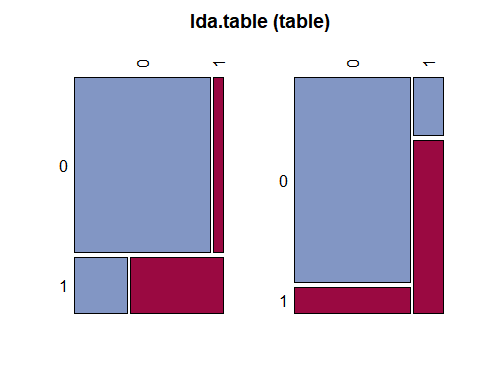
##   
## True infection rate:

mean(AIDS$infected != 0)

## [1] 0.2435718

lda.table <- table(AIDS$infected, lda.pred$class)  
  
library(DescTools)  
Desc(lda.table, digits = 2)

## ------------------------------------------------------------------------------   
## lda.table (table)  
##   
## Summary:   
## n: 2'139, rows: 2, columns: 2  
##   
## Pearson's Chi-squared test (cont. adj):  
## X-squared = 772.54, df = 1, p-value < 2.2e-16  
## Fisher's exact test p-value < 2.2e-16  
## McNemar's chi-squared = 19.763, df = 1, p-value = 8.765e-06  
##   
## estimate lwr.ci upr.ci'  
##   
## odds ratio 23.849 18.338 31.016  
## rel. risk (col1) 2.568 2.290 2.879  
## rel. risk (col2) 0.108 0.089 0.130  
## 0.569 0.525 0.611  
##   
##   
## Contingency Coeff. 0.516  
## Cramer's V 0.602  
## Kendall Tau-b 0.602  
##   
##   
## 0 1 Sum  
##   
## 0 freq 1'507 111 1'618  
## perc 70.45% 5.19% 75.64%  
## p.row 93.14% 6.86% .  
## p.col 88.86% 25.06% .  
##   
## 1 freq 189 332 521  
## perc 8.84% 15.52% 24.36%  
## p.row 36.28% 63.72% .  
## p.col 11.14% 74.94% .  
##   
## Sum freq 1'696 443 2'139  
## perc 79.29% 20.71% 100.00%  
## p.row . . .  
## p.col . . .  
##   
##   
## ----------  
## ' 95% conf. level



R console에서 확인할 수 있는 혼동행렬 결과를 보면 비감염(0, 0) 예측 정확도가 70.5%, 감염(1, 1) 15.52%로 총합 86.02% 에 달했다. 이는 로지스틱 판별 분석보다 약간 낮은 정도이다. 하지만 감염율 예측에 한정하면 약간 더 나은 정확도를 보인다. 이제부터는 정확도를 높이는 데 집중하겠다.

## 이차 판별 분석

이차 판별 분석은 이차함수 곡선과 선형대수의 이차 형태를 응용하여 곡선의 판별함수를 생성하는 모형이다. 곡선을 사용하기 때문에 선형적인 판별 모형보다 더 성능이 좋을 수가 있다.

boxM(num.vars, AIDS$infected) -> AIDS.M  
AIDS.M

##   
## Box's M-test for Homogeneity of Covariance Matrices  
##   
## data: num.vars  
## Chi-Sq (approx.) = 190.58, df = 45, p-value < 2.2e-16

선형 판별 모형의 등분산성 가정이 기각되었으므로 이차 판별 분석을 하는 것은 타당하다.

library(caret)

## 필요한 패키지를 로딩중입니다: lattice

nearZeroVar(AIDS, saveMetrics = TRUE) -> nzv  
AIDS[, !nzv$nzv] -> AIDS\_filtered  
  
# Apply PCA to reduce dimensionality (optional step if needed)  
prcomp(num.vars, center = TRUE, scale. = TRUE) -> pca  
data.frame(pca$x) -> AIDS\_pca   
  
AIDS$infected -> AIDS\_pca$infected  
  
  
# Perform QDA  
AIDS\_pca |>  
 qda(infected ~ ., data = \_) -> qda.result  
qda.result

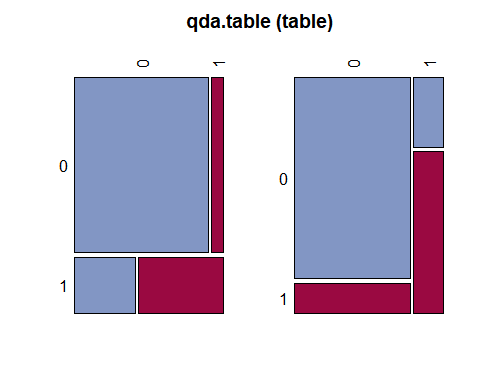
## Call:  
## qda(infected ~ ., data = AIDS\_pca)  
##   
## Prior probabilities of groups:  
## 0 1   
## 0.7564282 0.2435718   
##   
## Group means:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## 0 -0.2171578 0.2873816 -0.06081343 -0.03588718 0.02505281 -0.1743599  
## 1 0.6743979 -0.8924827 0.18886014 0.11145000 -0.07780315 0.5414863  
## PC7 PC8 PC9  
## 0 -0.04288246 -0.01630108 -0.006484794  
## 1 0.13317431 0.05062408 0.020138959

# Predict and create the confusion matrix  
predict(qda.result, AIDS\_pca) -> qda.pred  
table(AIDS\_pca$infected, qda.pred$class) -> qda.table  
  
# Display the results  
qda.table

##   
## 0 1  
## 0 1488 130  
## 1 218 303

Desc(qda.table, digits = 2)

## ------------------------------------------------------------------------------   
## qda.table (table)  
##   
## Summary:   
## n: 2'139, rows: 2, columns: 2  
##   
## Pearson's Chi-squared test (cont. adj):  
## X-squared = 610.14, df = 1, p-value < 2.2e-16  
## Fisher's exact test p-value < 2.2e-16  
## McNemar's chi-squared = 21.75, df = 1, p-value = 3.106e-06  
##   
## estimate lwr.ci upr.ci'  
##   
## odds ratio 15.909 12.392 20.425  
## rel. risk (col1) 2.198 1.984 2.435  
## rel. risk (col2) 0.138 0.115 0.165  
## 0.501 0.456 0.545  
##   
##   
## Contingency Coeff. 0.472  
## Cramer's V 0.535  
## Kendall Tau-b 0.535  
##   
##   
## 0 1 Sum  
##   
## 0 freq 1'488 130 1'618  
## perc 69.57% 6.08% 75.64%  
## p.row 91.97% 8.03% .  
## p.col 87.22% 30.02% .  
##   
## 1 freq 218 303 521  
## perc 10.19% 14.17% 24.36%  
## p.row 41.84% 58.16% .  
## p.col 12.78% 69.98% .  
##   
## Sum freq 1'706 433 2'139  
## perc 79.76% 20.24% 100.00%  
## p.row . . .  
## p.col . . .  
##   
##   
## ----------  
## ' 95% conf. level



오히려 정확도가 더 떨어졌다. 감염율 예측은 더 감소하였다. 오류 메시지에 따르면 변수 간 다중공선성이 있다고 한다.> AIDS |> + lda(infected ~ . , data = \_) -> lda.result lda.default(x, grouping, …)에서 경고가 발생했습니다 : 변수들이 공선형관계에 있습니다

이러한 경우 단순한 선형대수로, 특히 역행렬로 구할 수 있는 통계적 추정치들은 부정확할 수 있다. 이것이 더 어렵고 복잡한 머신러닝이 필요한 이유이다.

# 머신러닝

머신러닝은 설명보다는 예측에 중점을 두며, 대규모 데이터를 효율적으로 처리하면서도 충분히 유의미한 정보를 제공하기 위해 사용된다. 단점은 다변량 분석보다도 더 적절한 설명을 제공하기가 애매하다는 것이다. 그래서 그냥 예측 성능을 극대화하고 그 예측 결과를 그대로 제공하는 식의 접근이 많이 사용되고 있다. R에서 머신러닝은 *tidymodels* 패키지로 쉽게 할 수 있다.

## 모델 정의

tidymodels\_prefer()  
library(conflicted)  
  
# 디버그 ==> scales::alpha를 선호하도록 설정  
conflicts\_prefer(scales::alpha)

## [conflicted] Will prefer scales::alpha over any other package.

boost\_tree(learn\_rate = tune(), trees = tune(), min\_n = tune(), tree\_depth = tune()) %>%  
 set\_engine("xgboost") %>%set\_mode("classification") -> xgb.mod  
  
rand\_forest(trees = tune(), min\_n = tune()) %>%  
 set\_engine("ranger") %>% set\_mode("classification") -> rf.mod  
  
svm\_rbf(cost = tune(), rbf\_sigma = tune()) %>%  
 set\_engine("kernlab") %>% set\_mode("classification") -> svm.mod  
  
decision\_tree(cost\_complexity = tune(), tree\_depth = tune(), min\_n = tune()) %>%  
 set\_engine("rpart") %>% set\_mode("classification") -> dt.mod  
xgb.mod;rf.mod;svm.mod;dt.mod

## Boosted Tree Model Specification (classification)  
##   
## Main Arguments:  
## trees = tune()  
## min\_n = tune()  
## tree\_depth = tune()  
## learn\_rate = tune()  
##   
## Computational engine: xgboost

## Random Forest Model Specification (classification)  
##   
## Main Arguments:  
## trees = tune()  
## min\_n = tune()  
##   
## Computational engine: ranger

## Radial Basis Function Support Vector Machine Model Specification (classification)  
##   
## Main Arguments:  
## cost = tune()  
## rbf\_sigma = tune()  
##   
## Computational engine: kernlab

## Decision Tree Model Specification (classification)  
##   
## Main Arguments:  
## cost\_complexity = tune()  
## tree\_depth = tune()  
## min\_n = tune()  
##   
## Computational engine: rpart

대표적인 머신러닝 모델들인 XGBoost, RandomForest, RBF Kernel Support Vector Machine, Decision Tree 들을 사용할 것이다. 정답이 전혀 없는 머신러닝의 특성상 여러 개의 모델들을 비교하면서 가장 좋은 것을 찾는 것이 나쁘지 않는 접근법이 된다.

## 전처리

*tidymodels*에선 전처리를 “레시피”라고 한다. 레시피 객체에 공식과 데이터를 넣고, 몇 가지 전처리를 “step”들을 추가할 수 있다. 그 전에 훈련 데이터와 평가 데이터를 분할해야 성능을 제대로 평가할 수 있다.

AIDS %>% initial\_split(prop = 3/4, strata = infected) -> spliter  
spliter %>% training() -> aids.train  
spliter %>% testing() -> aids.test  
aids.train %>% skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 1603 |
| Number of columns | 23 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 14 |
| numeric | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| trt | 0 | 1 | FALSE | 4 | 3: 410, 0: 407, 1: 397, 2: 389 |
| hemo | 0 | 1 | FALSE | 2 | 0: 1480, 1: 123 |
| homo | 0 | 1 | FALSE | 2 | 1: 1077, 0: 526 |
| drugs | 0 | 1 | FALSE | 2 | 0: 1393, 1: 210 |
| oprior | 0 | 1 | FALSE | 2 | 0: 1576, 1: 27 |
| z30 | 0 | 1 | FALSE | 2 | 1: 898, 0: 705 |
| race | 0 | 1 | FALSE | 2 | 0: 1147, 1: 456 |
| gender | 0 | 1 | FALSE | 2 | 1: 1341, 0: 262 |
| str2 | 0 | 1 | FALSE | 2 | 1: 950, 0: 653 |
| strat | 0 | 1 | FALSE | 3 | 1: 653, 3: 646, 2: 304 |
| symptom | 0 | 1 | FALSE | 2 | 0: 1313, 1: 290 |
| treat | 0 | 1 | FALSE | 2 | 1: 1196, 0: 407 |
| offtrt | 0 | 1 | FALSE | 2 | 0: 1028, 1: 575 |
| infected | 0 | 1 | FALSE | 2 | 0: 1213, 1: 390 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| time | 0 | 1 | 881.64 | 289.77 | 14.00 | 732.0 | 997.00 | 1091.0 | 1231.00 | ▁▂▂▃▇ |
| age | 0 | 1 | 35.33 | 8.78 | 12.00 | 29.0 | 34.00 | 40.0 | 70.00 | ▁▇▆▂▁ |
| wtkg | 0 | 1 | 75.26 | 13.23 | 32.66 | 66.8 | 74.39 | 82.3 | 159.94 | ▁▇▂▁▁ |
| karnof | 0 | 1 | 95.37 | 5.88 | 70.00 | 90.0 | 100.00 | 100.0 | 100.00 | ▁▁▁▅▇ |
| preanti | 0 | 1 | 382.90 | 464.20 | 0.00 | 0.0 | 160.00 | 747.0 | 2851.00 | ▇▃▁▁▁ |
| cd40 | 0 | 1 | 349.02 | 120.36 | 0.00 | 261.0 | 337.00 | 421.0 | 1199.00 | ▂▇▂▁▁ |
| cd420 | 0 | 1 | 370.86 | 144.12 | 50.00 | 269.5 | 353.00 | 460.0 | 1119.00 | ▃▇▃▁▁ |
| cd80 | 0 | 1 | 989.17 | 488.77 | 40.00 | 650.0 | 892.00 | 1213.0 | 5011.00 | ▇▅▁▁▁ |
| cd820 | 0 | 1 | 935.49 | 445.35 | 131.00 | 630.0 | 871.00 | 1140.0 | 6035.00 | ▇▁▁▁▁ |

aids.test %>% skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 536 |
| Number of columns | 23 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 14 |
| numeric | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| trt | 0 | 1 | FALSE | 4 | 3: 151, 2: 135, 0: 125, 1: 125 |
| hemo | 0 | 1 | FALSE | 2 | 0: 479, 1: 57 |
| homo | 0 | 1 | FALSE | 2 | 1: 337, 0: 199 |
| drugs | 0 | 1 | FALSE | 2 | 0: 465, 1: 71 |
| oprior | 0 | 1 | FALSE | 2 | 0: 516, 1: 20 |
| z30 | 0 | 1 | FALSE | 2 | 1: 279, 0: 257 |
| race | 0 | 1 | FALSE | 2 | 0: 375, 1: 161 |
| gender | 0 | 1 | FALSE | 2 | 1: 430, 0: 106 |
| str2 | 0 | 1 | FALSE | 2 | 1: 303, 0: 233 |
| strat | 0 | 1 | FALSE | 3 | 1: 233, 3: 197, 2: 106 |
| symptom | 0 | 1 | FALSE | 2 | 0: 456, 1: 80 |
| treat | 0 | 1 | FALSE | 2 | 1: 411, 0: 125 |
| offtrt | 0 | 1 | FALSE | 2 | 0: 335, 1: 201 |
| infected | 0 | 1 | FALSE | 2 | 0: 405, 1: 131 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| time | 0 | 1 | 871.51 | 299.78 | 50 | 654.25 | 998.50 | 1091.00 | 1231.00 | ▁▂▂▃▇ |
| age | 0 | 1 | 35.01 | 8.48 | 12 | 29.00 | 34.00 | 40.00 | 65.00 | ▁▇▇▂▁ |
| wtkg | 0 | 1 | 74.73 | 13.37 | 31 | 66.28 | 74.25 | 82.80 | 120.66 | ▁▃▇▂▁ |
| karnof | 0 | 1 | 95.67 | 5.95 | 70 | 90.00 | 100.00 | 100.00 | 100.00 | ▁▁▁▅▇ |
| preanti | 0 | 1 | 368.04 | 482.01 | 0 | 0.00 | 100.50 | 718.25 | 2500.00 | ▇▃▁▁▁ |
| cd40 | 0 | 1 | 354.93 | 113.05 | 99 | 265.75 | 350.00 | 431.00 | 770.00 | ▃▇▇▂▁ |
| cd420 | 0 | 1 | 372.63 | 146.28 | 49 | 268.50 | 357.00 | 461.00 | 1100.00 | ▃▇▃▁▁ |
| cd80 | 0 | 1 | 979.03 | 453.96 | 105 | 670.00 | 901.50 | 1181.50 | 4255.00 | ▇▆▁▁▁ |
| cd820 | 0 | 1 | 935.00 | 444.26 | 124 | 632.75 | 839.00 | 1162.25 | 3407.00 | ▇▇▂▁▁ |

요약통계표를 보니 분할은 잘 된 것 같다. 이제 레시피를 작성하겠다.

recipe(infected ~ ., data = aids.train) %>%  
 step\_dummy(all\_nominal\_predictors()) %>%  
 step\_normalize(all\_numeric\_predictors()) %>%  
 step\_corr(all\_numeric\_predictors()) %>%  
 step\_zv(all\_numeric\_predictors()) -> aids.base.recipe  
  
recipe(infected ~ ., data = aids.train) %>%  
 step\_dummy(all\_nominal\_predictors()) %>%  
 step\_pca(all\_numeric\_predictors(), num\_comp = 7) %>%  
 step\_normalize(all\_numeric\_predictors()) %>%  
 step\_corr(all\_numeric\_predictors()) %>%  
 step\_zv(all\_numeric\_predictors()) -> aids.pca.recipe  
aids.base.recipe;aids.pca.recipe

##

## ── Recipe ──────────────────────────────────────────────────────────────────────

##

## ── Inputs

## Number of variables by role

## outcome: 1  
## predictor: 22

##

## ── Operations

## • Dummy variables from: all\_nominal\_predictors()

## • Centering and scaling for: all\_numeric\_predictors()

## • Correlation filter on: all\_numeric\_predictors()

## • Zero variance filter on: all\_numeric\_predictors()

##

## ── Recipe ──────────────────────────────────────────────────────────────────────

##

## ── Inputs

## Number of variables by role

## outcome: 1  
## predictor: 22

##

## ── Operations

## • Dummy variables from: all\_nominal\_predictors()

## • PCA extraction with: all\_numeric\_predictors()

## • Centering and scaling for: all\_numeric\_predictors()

## • Correlation filter on: all\_numeric\_predictors()

## • Zero variance filter on: all\_numeric\_predictors()

전처리는 숫자형 변수들에게만 하였다. 기본적으로 정규화를 하고, 이후에 상관관계가 강한 변수를 제거하고, 분산이 0이 되어버린 변수도 제거한다. 또 다른 하나의 전처리 과정은 pca를 먼저 하고 그것들을 적용한다. 앞선 주성분 요인분석에서 최적의 주성분 개수가 7개로 드러났으므로 요인의 개수를 지정하는 인수 *num\_comp*에 7을 지정했다.

## 워크플로우 작성

*tidymodels* 패키지의 특장점은 레시피와 모델을 손쉽게 하나의 워크플로우로 묶을 수 있고, 심지어는 그 워크플로우들 자체도 묶을 수 있다는 것이다. 그러한 작업은 *workflow\_set()*으로 간단하게 할 수 있다.

workflow\_set(  
 preproc = list(Base = aids.base.recipe,  
 PCA = aids.pca.recipe),  
 models = list(XGB = xgb.mod,  
 RandForest = rf.mod,  
 SVM = svm.mod,  
 DecisionTree = dt.mod)  
) -> aids.wflow  
aids.wflow

## # A workflow set/tibble: 8 × 4  
## wflow\_id info option result   
## <chr> <list> <list> <list>   
## 1 Base\_XGB <tibble [1 × 4]> <opts[0]> <list [0]>  
## 2 Base\_RandForest <tibble [1 × 4]> <opts[0]> <list [0]>  
## 3 Base\_SVM <tibble [1 × 4]> <opts[0]> <list [0]>  
## 4 Base\_DecisionTree <tibble [1 × 4]> <opts[0]> <list [0]>  
## 5 PCA\_XGB <tibble [1 × 4]> <opts[0]> <list [0]>  
## 6 PCA\_RandForest <tibble [1 × 4]> <opts[0]> <list [0]>  
## 7 PCA\_SVM <tibble [1 × 4]> <opts[0]> <list [0]>  
## 8 PCA\_DecisionTree <tibble [1 × 4]> <opts[0]> <list [0]>

이렇게 간단하게 여러 개의 워크플로우 객체들이 묶여서 하나의 객체가 되었다. 이제 이 객체에 *workflow\_map()*함수를 적용하면 모델을 전부 훈련시킬 수 있다.

## 모델 훈련

다양한 하이퍼파라미터를 수동으로 최적화하는 것보다 좋은 방법은 반복적인 교차 검증을 통해 가능한 경우의 수 중에서 가장 우수한 하이퍼파라미터 조합을 컴퓨터가 알아서 찾아내고 저장하게 하는 것이다. 그러고 나면 사람이 할 일은 그냥 정렬을 한 다음 가장 위에 있는 걸 선택하는 것이다. *workflow\_map()*함수는 인수로 자동 하이퍼파라미터 튜닝 기법을 내장하고 있어서 손쉽게 할 수 있다. 다만 *vfold\_cv()*로 생성된 교차 검증용 객체도 함께 입력해야 오류 없이 훈련할 수 있다.

vfold\_cv(aids.train, v = 5, strata = infected) -> vfold.aids  
  
aids.wflow %>%  
 workflow\_map(  
 fn = "tune\_bayes",  
 resamples = vfold.aids,  
 metrics = metric\_set(accuracy, roc\_auc),  
 verbose = TRUE  
 ) -> aids.wflow.res

## i 1 of 8 tuning: Base\_XGB

## ! The Gaussian process model is being fit using 4 features but only has 5  
## data points to do so. This may cause errors or a poor model fit.

## ! No improvement for 10 iterations; returning current results.

## ✔ 1 of 8 tuning: Base\_XGB (3m 22.3s)

## i 2 of 8 tuning: Base\_RandForest

## ! No improvement for 10 iterations; returning current results.

## ✔ 2 of 8 tuning: Base\_RandForest (1m 8.1s)

## i 3 of 8 tuning: Base\_SVM

## ! All of the accuracy values were identical. The Gaussian process model  
## cannot be fit to the data. Try expanding the range of the tuning  
## parameters.

## → A | error: Infinite values of the Deviance Function,   
## unable to find optimum parameters   
##

## There were issues with some computations A: x1 ✖ Optimization stopped prematurely; returning current results.  
## There were issues with some computations A: x1There were issues with some computations A: x1  
## ✔ 3 of 8 tuning: Base\_SVM (14.2s)  
## i 4 of 8 tuning: Base\_DecisionTree  
## ✔ 4 of 8 tuning: Base\_DecisionTree (41.2s)  
## i 5 of 8 tuning: PCA\_XGB  
## ! The Gaussian process model is being fit using 4 features but only has 5  
## data points to do so. This may cause errors or a poor model fit.  
## ✔ 5 of 8 tuning: PCA\_XGB (2m 47.7s)  
## i 6 of 8 tuning: PCA\_RandForest  
## ✔ 6 of 8 tuning: PCA\_RandForest (1m 15s)  
## i 7 of 8 tuning: PCA\_SVM  
## ! All of the accuracy values were identical. The Gaussian process model  
## cannot be fit to the data. Try expanding the range of the tuning  
## parameters.  
## → A | error: Infinite values of the Deviance Function,   
## unable to find optimum parameters   
##   
## There were issues with some computations A: x1 ✖ Optimization stopped prematurely; returning current results.  
## There were issues with some computations A: x1There were issues with some computations A: x1  
## ✔ 7 of 8 tuning: PCA\_SVM (11.8s)  
## i 8 of 8 tuning: PCA\_DecisionTree  
## ✔ 8 of 8 tuning: PCA\_DecisionTree (36.2s)

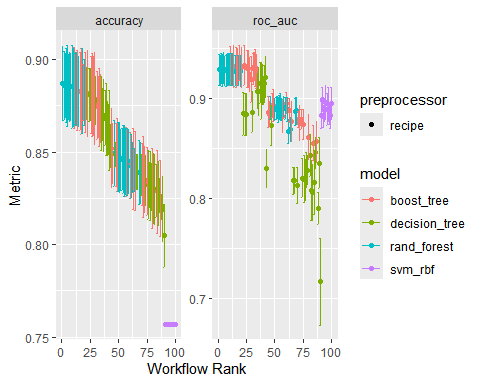
aids.wflow.res

## # A workflow set/tibble: 8 × 4  
## wflow\_id info option result   
## <chr> <list> <list> <list>   
## 1 Base\_XGB <tibble [1 × 4]> <opts[2]> <tune[+]>  
## 2 Base\_RandForest <tibble [1 × 4]> <opts[2]> <tune[+]>  
## 3 Base\_SVM <tibble [1 × 4]> <opts[2]> <tune[+]>  
## 4 Base\_DecisionTree <tibble [1 × 4]> <opts[2]> <tune[+]>  
## 5 PCA\_XGB <tibble [1 × 4]> <opts[2]> <tune[+]>  
## 6 PCA\_RandForest <tibble [1 × 4]> <opts[2]> <tune[+]>  
## 7 PCA\_SVM <tibble [1 × 4]> <opts[2]> <tune[+]>  
## 8 PCA\_DecisionTree <tibble [1 × 4]> <opts[2]> <tune[+]>

## 최적 모델 선택

*autoplot()*으로 모델 평가 지표를 시각화하고, *rank\_results()*로 최고의 지표를 가진 워크플로우 객체를 확인할 수 있다.

aids.wflow.res %>% autoplot()



aids.wflow.res %>%  
 rank\_results()

## # A tibble: 200 × 9  
## wflow\_id .config .metric mean std\_err n preprocessor model rank  
## <chr> <chr> <chr> <dbl> <dbl> <int> <chr> <chr> <int>  
## 1 Base\_RandForest Preproc… accura… 0.887 0.0122 5 recipe rand… 1  
## 2 Base\_RandForest Preproc… roc\_auc 0.929 0.00948 5 recipe rand… 1  
## 3 Base\_RandForest Iter4 accura… 0.886 0.0113 5 recipe rand… 2  
## 4 Base\_RandForest Iter4 roc\_auc 0.928 0.00927 5 recipe rand… 2  
## 5 Base\_RandForest Iter1 accura… 0.886 0.0105 5 recipe rand… 3  
## 6 Base\_RandForest Iter1 roc\_auc 0.930 0.00939 5 recipe rand… 3  
## 7 Base\_RandForest Iter2 accura… 0.886 0.0121 5 recipe rand… 4  
## 8 Base\_RandForest Iter2 roc\_auc 0.928 0.00980 5 recipe rand… 4  
## 9 Base\_RandForest Preproc… accura… 0.886 0.0133 5 recipe rand… 5  
## 10 Base\_RandForest Preproc… roc\_auc 0.930 0.00931 5 recipe rand… 5  
## # ℹ 190 more rows

PCA를 통한 차원 축소를 하지 않은 레시피를 사용한 Base\_XGB 워크플로우, 즉 가변수 변환된 범주형 변수들과 정규화된 숫자형 변수들 중, 강력한 상관관계를 가진 변수와 분산이 0인 변수를 제외한 전처리 데이터로 훈련 및 예측을 실시하는 XGBoost 분류기가 최고의 모델로 밝혀졌다.

## 최종 평가

평가 데이터 *aids.test*로 최종적인 평가 지표들을 확인한다. 먼저 *workflow\_map*의 결과에서 앞서 확인한 **wflow\_id**상의 이름을 *extract\_workflow\_set\_result()*에 입력해 결과 테이블을 추출하고, 거기서 *select\_best(metric=)*으로 가장 좋은 워크플로우의 하이퍼파라미터들을 추출한 다음, 그걸 빈 워크풀로우 세트에서 *extract\_workflow(“워크플로우 이름”)*로 추출한 워크플로우에 *finalaize\_workflow()*를 적용할 때 입력한다. 그리고 훈련 데이터에 *fit()*으로 적합시키고 *predict()*로 평가 데이터를 예측하고, *reframe()*으로 예측값과 실제값을 하나의 테이블로 합친 다음 *metrics(estimate = 예측값 컬럼명, truth = 평가 데이터 라벨)*에 입력해 최종적인 평가 지표들을 얻는다.

aids.wflow.res %>%  
 extract\_workflow\_set\_result(id = "Base\_XGB") %>%  
 select\_best(metric = "accuracy") -> xgb.params  
  
aids.wflow %>%   
 extract\_workflow("Base\_XGB") %>%  
 finalize\_workflow(xgb.params) %>%  
 fit(aids.train) %>%  
 predict(new\_data = aids.test) %>%  
 reframe(pred = .pred\_class, true = aids.test$infected) %>%  
 metrics(estimate = pred, truth = true)

## # A tibble: 2 × 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 accuracy binary 0.896  
## 2 kap binary 0.711

최종 정확도가 89.55%로 크게 향상되었다.