변수 제거

1. 주성분 분석

2. 0에 가까운 분산을 가지는 변수 제거

분산이 0에 가까운 변수는 제거해도 큰 영향이 없음.
nearZeroVar()함수
https://www.rdocumentation.org/packages/caret/versions/6.0-86/topics/nearZeroVar

where
'saveMetrics=FALSE'속성: 예측변수의 컬럼위치에 해당하는 정수 벡터
'saveMetrics=TRUE'속성: 컬럼을 가지는 데이터프레임
freqRatio: 가장 큰 공통값 대비 두번째 큰 공통값의 빈도의 비율
percentUnique: 데이터 전체로 부터 고유 데이터의 비율
zeroVar: 예측변수가 오직 한개의 특이값을 갖는지 여부에 대한 논리 벡터
nzv: 예측변수가 0에 가까운 분산에측 변수인지 여부에 대한 논리 벡터

실습.

```
========================
install.packages("caret")
library(caret)
install.packages("mlbench")
library(mlbench)
nearZeroVar(iris, saveMetrics=TRUE)

data(Soybean)
head(Soybean)
# 0에 가까운 분산을 가지는 변수의 존재 여부 확인
nearZeroVar(Soybean, saveMetrics=TRUE)

======================
```

```
> nearZeroVar(Soybean, saveMetrics=TRUE)
                 freqRatio percentUnique zeroVar   nzv
class             1.010989      2.7818448   FALSE FALSE
date              1.137405      1.0248902   FALSE FALSE
plant.stand       1.208191      0.2928258   FALSE FALSE
precip            4.098214      0.4392387   FALSE FALSE
temp              1.879397      0.4392387   FALSE FALSE
hail              3.425197      0.2928258   FALSE FALSE
crop.hist         1.004587      0.5856515   FALSE FALSE
area.dam          1.213904      0.5856515   FALSE FALSE
sever             1.651282      0.4392387   FALSE FALSE
seed.tmt          1.373874      0.4392387   FALSE FALSE
germ              1.103627      0.4392387   FALSE FALSE
plant.growth      1.951327      0.2928258   FALSE FALSE
leaves            7.870130      0.2928258   FALSE FALSE
leaf.halo         1.547511      0.4392387   FALSE FALSE
leaf.marg         1.615385      0.4392387   FALSE FALSE
leaf.size         1.479638      0.4392387   FALSE FALSE
leaf.shread       5.072917      0.2928258   FALSE FALSE
leaf.malf        12.311111      0.2928258   FALSE FALSE
leaf.mild        26.750000      0.4392387   FALSE  TRUE
stem              1.253378      0.2928258   FALSE FALSE
lodging          12.380952      0.2928258   FALSE FALSE
stem.cankers      1.984293      0.5856515   FALSE FALSE
canker.lesion     1.807910      0.5856515   FALSE FALSE
fruiting.bodies   4.548077      0.2928258   FALSE FALSE
ext.decay         3.681481      0.4392387   FALSE FALSE
mycelium        106.500000      0.2928258   FALSE  TRUE
int.discolor     13.204545      0.4392387   FALSE FALSE
sclerotia        31.250000      0.2928258   FALSE  TRUE
fruit.pods        3.130769      0.5856515   FALSE FALSE
fruit.spots       3.450000      0.5856515   FALSE FALSE
seed              4.139130      0.2928258   FALSE FALSE
mold.growth       7.820896      0.2928258   FALSE FALSE
seed.discolor     8.015625      0.2928258   FALSE FALSE
seed.size         9.016949      0.2928258   FALSE FALSE
shriveling       14.184211      0.2928258   FALSE FALSE
roots             6.406977      0.4392387   FALSE FALSE
>
```

nzv = 'TRUE' 인 leaf.mild, mycelium, sclerotia 변수를 제거 해도 큰 영향이 없다.

3. 상관관계가 높은 변수 제거

상관관계가 높은 컬럼을 제외

findCorrelation()함수
https://www.rdocumentation.org/packages/caret/versions/6.0-88/topics/findCorrelation

실습.

```
============
library(caret)

library(mlbench)
data(Vehicle)
head(Vehicle)

# 상관관계 높은 열 선정
findCorrelation(cor(subset(Vehicle, select=-c(Class))))

# 상관관계가 높은 열끼리 상관관계 확인
cor(subset(Vehicle, select=-c(Class))) [c(3,8,11,7,9,2), c(3,8,11,7,9,2)]

# 상관관계 높은 열 제거
Cor_Vehicle <- Vehicle[,-c(3,8,11,7,9,2)]

findCorrelation(cor(subset(Cor_Vehicle, select=-c(Class))))

head(Cor_Vehicle)

======================
```

```
>
>
> # 상관관계 높은 열 선정
> findCorrelation(cor(subset(Vehicle, select=-c(Class))))
[1]  3  8 11  7  9  2
> # 상관관계가 높은 열끼리 상관관계 확인
> cor(subset(Vehicle, select=-c(Class))) [c(3,8,11,7,9,2), c(3,8,11,7,9,2)]
                D.Circ      Elong Sc.Var.Maxis    Scat.Ra Pr.Axis.Rect
D.Circ       1.0000000 -0.9123072    0.8644323  0.9072801    0.8953261
Elong       -0.9123072  1.0000000   -0.9383919 -0.9733853   -0.9505124
Sc.Var.Maxis 0.8644323 -0.9383919    1.0000000  0.9518621    0.9382664
Scat.Ra      0.9072801 -0.9733853    0.9518621  1.0000000    0.9920883
Pr.Axis.Rect 0.8953261 -0.9505124    0.9382664  0.9920883    1.0000000
Circ         0.7984920 -0.8287548    0.8084963  0.8603671    0.8579253
                 Circ
D.Circ       0.7984920
Elong       -0.8287548
Sc.Var.Maxis 0.8084963
Scat.Ra      0.8603671
Pr.Axis.Rect 0.8579253
Circ         1.0000000
>
```

```
> # 상관관계 높은 열 제거
> Cor_Vehicle <- Vehicle[,-c(3,8,11,7,9,2)]
>
> findCorrelation(cor(subset(Cor_Vehicle, select=-c(Class))))
integer(0)
>
```

4. 카이 제곱 검정을 통한 중요 변수 선발

카이제곱검정을 실행하여 중요 변수 선발

실습.

```
==================
install.packages("FSelector")
library(FSelector)
library(mlbench)

data(Vehicle)

#카이 제곱 검정으로 변수들의 중요성 평가
(cs <- chi.squared(Class ~., data=Vehicle))

#변수 중에서 중요한 5개 선별
cutoff.k(cs,5)

====================
```

```
>
> #카이 제곱 검정으로 변수들의 중요성 평가
> (cs <- chi.squared(Class ~., data=Vehicle))
             attr_importance
Comp               0.3043172
Circ               0.2974762
D.Circ             0.3587826
Rad.Ra             0.3509038
Pr.Axis.Ra         0.2264652
Max.L.Ra           0.3234535
Scat.Ra            0.4653985
Elong              0.4556748
Pr.Axis.Rect       0.4475087
Max.L.Rect         0.3059760
Sc.Var.Maxis       0.4338378
Sc.Var.maxis       0.4921648
Ra.Gyr             0.2940064
Skew.Maxis         0.3087694
Skew.maxis         0.2470216
Kurt.maxis         0.3338930
Kurt.Maxis         0.2732117
Holl.Ra            0.3886266
> #변수 중에서 중요한 5개 선별
> cutoff.k(cs,5)
[1] "Sc.Var.maxis" "Scat.Ra"      "Elong"        "Pr.Axis.Rect" "Sc.Var.Maxis"
>
```