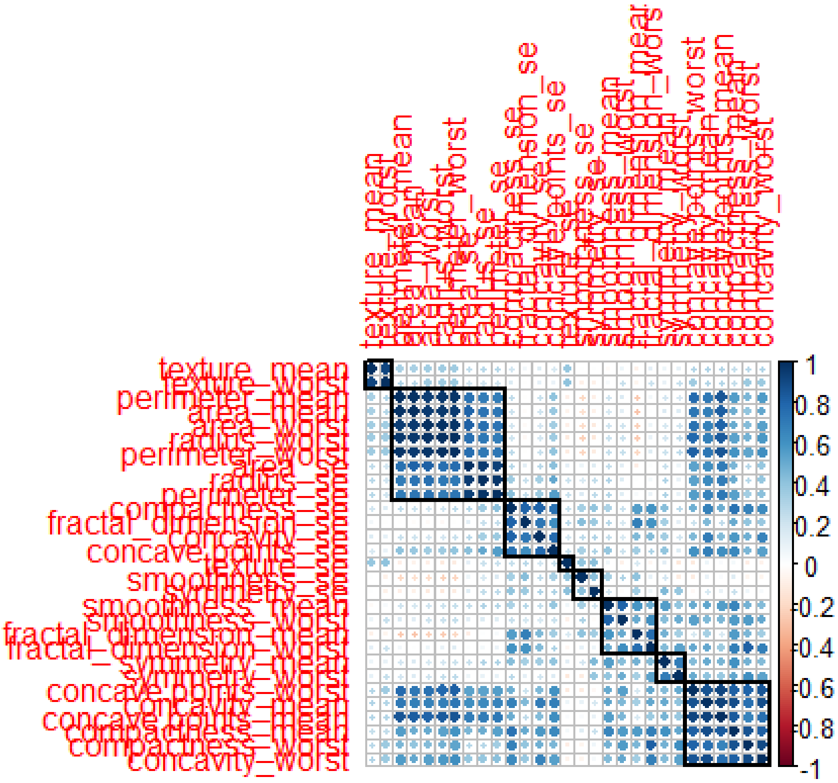
stProject - Data Mining

```
setwd("D:/Grad Study/Data Mining/Project/")
Cancer <- read.csv("Breast Cancer.csv")
```

```
str(Cancer)
```

```
## 'data.frame':    569 obs. of  33 variables:
##  $ id                     : int  842302 842517 84300903 84348301 84358402
843786 844359 84458202 844981 84501001 ...
##  $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2
2 ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149
...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511
...
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
##  $ X                      : logi  NA NA NA NA NA NA ...
```

```
Cancer$id <- NULL
Cancer$X <- NULL
anyNA(Cancer)

## [1] FALSE

Cancer1 <- Cancer[,c(-1:-2)]
correlations <- cor(Cancer1)
dim(correlations)

## [1] 29 29

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.5.3

## corrplot 0.84 loaded

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

corrplot(correlations, order = "hclust", tl.cex = 1, addrect = 8)
```



```
highCorr <- findCorrelation(correlations, cutoff = 0.85)
length(highCorr)
```

```
## [1] 12
```

```
filteredCancer <- Cancer[,-highCorr]
```

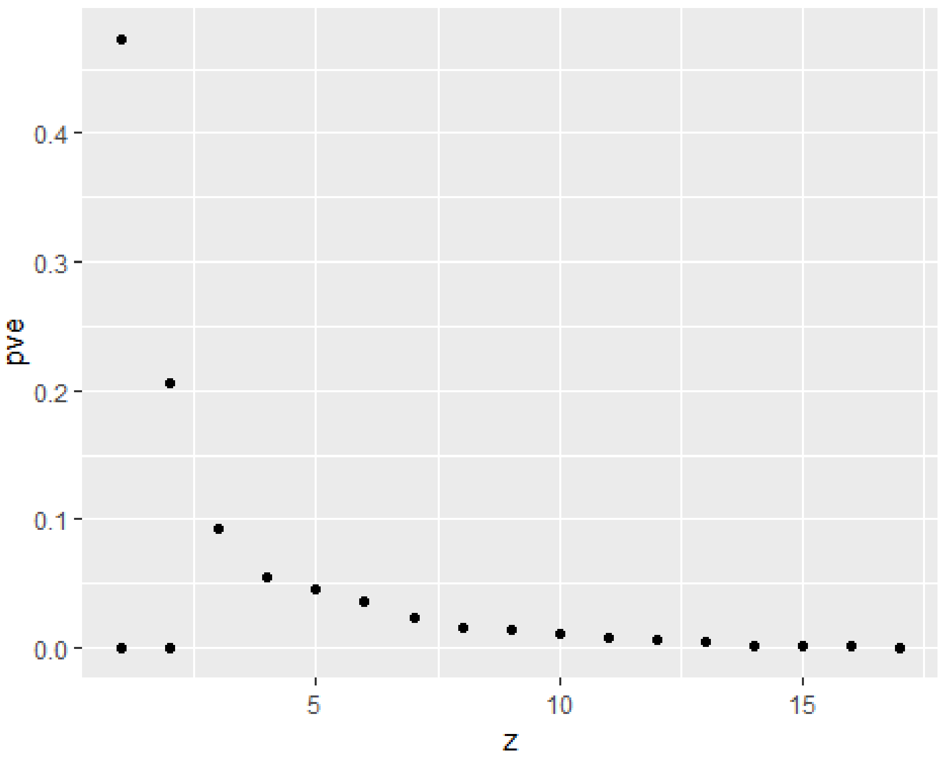
```
Cancer.trans <- preProcess(filteredCancer, method = c("BoxCox", "center",
"scale"))
Cancer.transformed <- predict(Cancer.trans, filteredCancer)
head(Cancer.transformed[,1:4])
```

```
##   texture_mean perimeter_mean concavity_mean concave.points_mean
## 1   -2.6966342      1.2560773     2.65054179           2.5302489
## 2   -0.2615935      1.5213622    -0.02382489           0.5476623
## 3    0.5484335      1.4483646     1.36227979           2.0354398
## 4    0.3590997     -0.5111072     1.91421287           1.4504311
## 5   -1.2329217      1.5751647     1.36980615           1.4272370
## 6   -0.8225400     -0.2467828     0.86554001           0.8239307
```

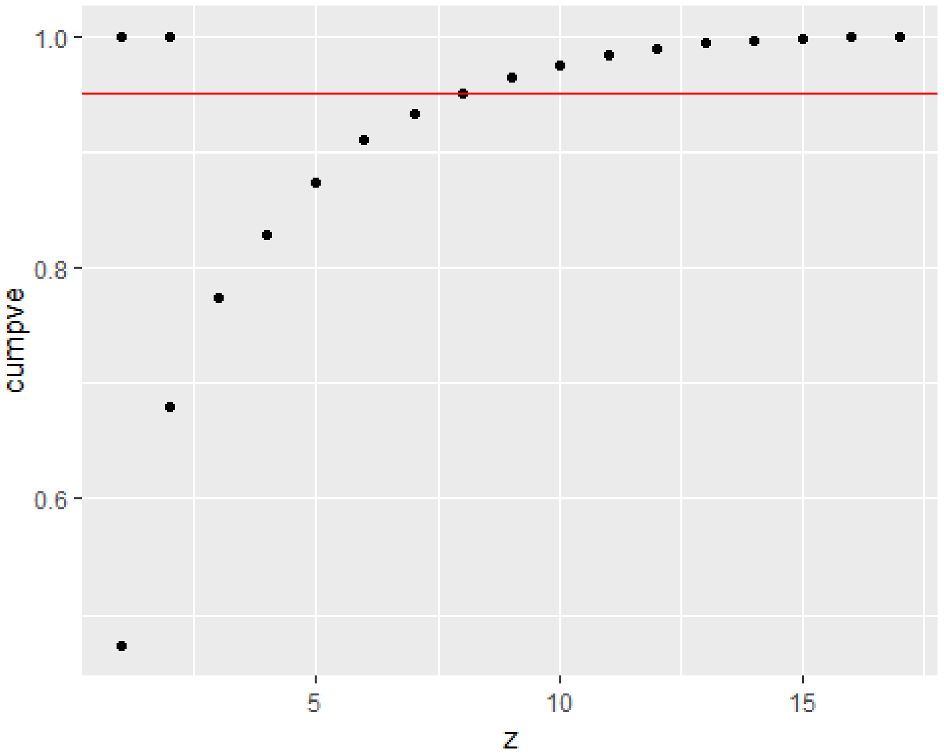
```
segmentation <- Cancer[,2]
pca.out <- prcomp(Cancer.transformed)
pca.var = pca.out$sdev^2
pve = pca.var/sum(pca.var)
z= seq(1,17)
cumpve = cumsum(pve)
pve.table = as.data.frame(cbind(z,pve, cumpve))
```

```
## Warning in cbind(z, pve, cumpve): number of rows of result is not a
## multiple of vector length (arg 1)
```

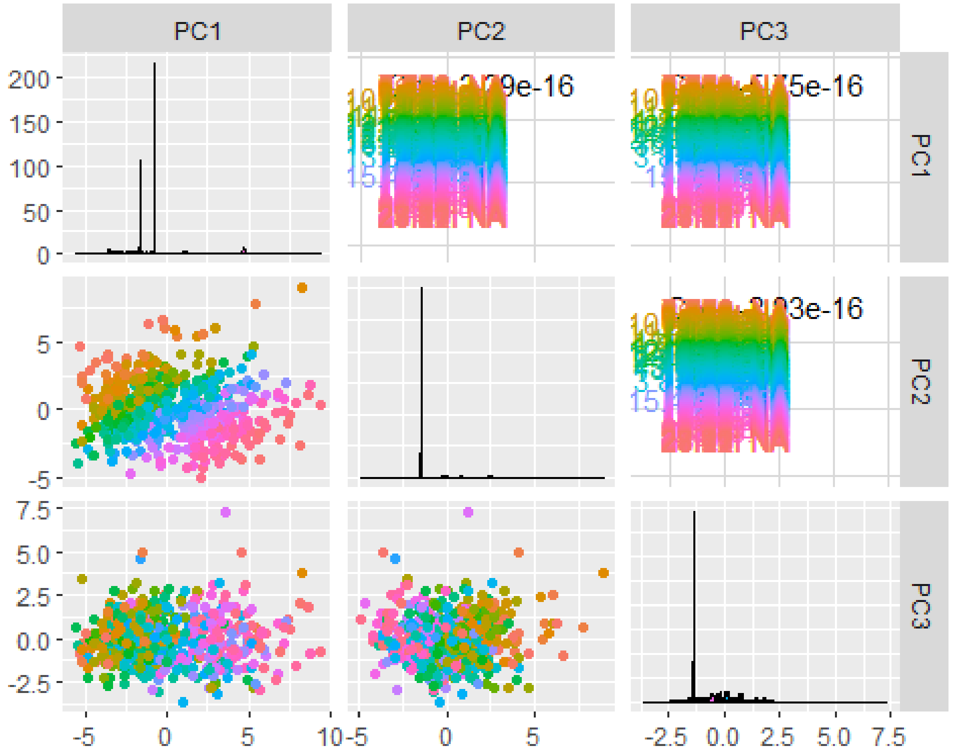
```
ggplot(pve.table, aes(x=z, y=pve))+ geom_point()
```



```r
ggplot(pve.table, aes(x=z, y=cumpve))+ geom_point() + geom_abline(intercept =
0.95, slope = 0, color = "red")
```

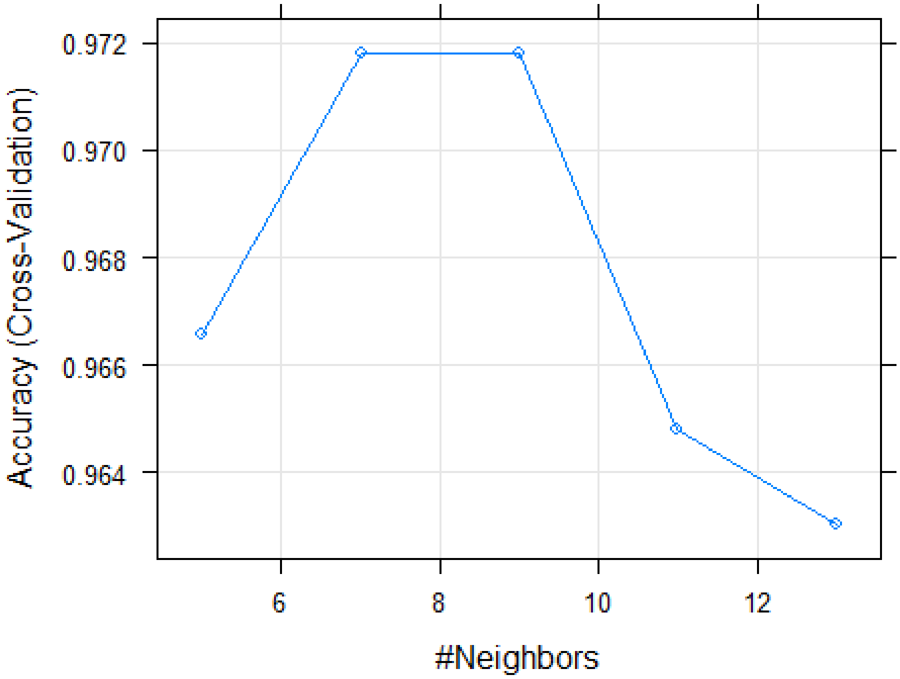


```r
library(GGally)

## Warning: package 'GGally' was built under R version 3.5.3

require(GGally)
PCs <- as.data.frame(cbind(segmentation, pca.out$x))
PCs$segmentation <- as.factor(PCs$segmentation)
ggpairs(data = PCs, columns = 2:4, ggplot2::aes(color = segmentation))
```



```r
library(ggplot2)
library(lattice)
library(caret)
set.seed(1)

DataPart <- createDataPartition(Cancer$diagnosis, p=0.8, list = F)
Train <- Cancer[DataPart,]
Test <- Cancer[-DataPart,]

set.seed(999)
ctrl <- trainControl(method = "cv", number = 5)

knn_c <- train(diagnosis~., data = Cancer, method = "knn", trControl = ctrl,
preProcess = c("center", "scale"), tuneLength = 5)
knn_c$results

##    k  Accuracy     Kappa AccuracySD    KappaSD
## 1  5 0.9665577 0.9274437 0.02005800 0.04350438
## 2  7 0.9718209 0.9389410 0.01580707 0.03430260
## 3  9 0.9718209 0.9389410 0.01580707 0.03430260
## 4 11 0.9647875 0.9233960 0.01769959 0.03867490
## 5 13 0.9630176 0.9193134 0.02285011 0.05014680
```

```
plot(knn_c)
```



```
Train_Scaled <- scale(Train[,-1], center = T, scale = T)
Test_Scaled <- scale(Test[,-1], center = T, scale = T)
library(class)
knn <- knn(train = Train_Scaled, test = Test_Scaled, cl=Train$diagnosis, k =
5)

mean(knn ==Test$diagnosis)

## [1] 0.9646018

summary(Cancer)

##  diagnosis  radius_mean      texture_mean   perimeter_mean
##  B:357     Min.   : 6.981   Min.   : 9.71   Min.   : 43.79
##  M:212     1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17
##            Median :13.370   Median :18.84   Median : 86.24
##            Mean   :14.127   Mean   :19.29   Mean   : 91.97
##            3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10
##            Max.   :28.110   Max.   :39.28   Max.   :188.50
##    area_mean      smoothness_mean   compactness_mean  concavity_mean
##  Min.   : 143.5   Min.   :0.05263   Min.   :0.01938   Min.   :0.00000
##  1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956
##  Median : 551.1   Median :0.09587   Median :0.09263   Median :0.06154
##  Mean   : 654.9   Mean   :0.09636   Mean   :0.10434   Mean   :0.08880
##  3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070
##  Max.   :2501.0   Max.   :0.16340   Max.   :0.34540   Max.   :0.42680
##  concave.points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##   radius_se         texture_se      perimeter_se       area_se
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509
##  Median :0.006380   Median :0.020450   Median :0.02589
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600
##  concave.points_se   symmetry_se       fractal_dimension_se
##  Min.   :0.000000   Min.   :0.007882   Min.   :0.0008948
##  1st Qu.:0.007638   1st Qu.:0.015160   1st Qu.:0.0022480
##  Median :0.010930   Median :0.018730   Median :0.0031870
##  Mean   :0.011796   Mean   :0.020542   Mean   :0.0037949
##  3rd Qu.:0.014710   3rd Qu.:0.023480   3rd Qu.:0.0045580
##  Max.   :0.052790   Max.   :0.078950   Max.   :0.0298400
##   radius_worst   texture_worst   perimeter_worst    area_worst
##  Min.   : 7.93   Min.   :12.02   Min.   : 50.41   Min.   : 185.2
##  1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11   1st Qu.: 515.3
##  Median :14.97   Median :25.41   Median : 97.66   Median : 686.5
##  Mean   :16.27   Mean   :25.68   Mean   :107.26   Mean   : 880.6
##  3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40   3rd Qu.:1084.0
##  Max.   :36.04   Max.   :49.54   Max.   :251.20   Max.   :4254.0
##  smoothness_worst  compactness_worst concavity_worst  concave.points_worst
##  Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493
##  Median :0.13130   Median :0.21190   Median :0.2267   Median :0.09993
##  Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean   :0.11461
##  3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140
##  Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.   :0.29100
##  symmetry_worst   fractal_dimension_worst
##  Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2822   Median :0.08004
##  Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :0.6638   Max.   :0.20750
```

```
summary(knn)
```

```
##  B  M
## 73 40
```

```
Pred1 <- train(diagnosis~., data = Cancer, method = "glm", trControl = ctrl,
tuneLength = 20)
```

```
Pred1$results
```

```
##   parameter  Accuracy     Kappa AccuracySD    KappaSD
## 1      none 0.9507685 0.8956496 0.01828296 0.03841739
```

```
library(ROCR)
```

```
n <- dim(Cancer)[1]
p <- 5
nsim <- round(n/5,0)
Pred_p <- predict(Pred1, Cancer, type = "prob")
Score <- prediction(Pred_p$B, Cancer$diagnosis)
Roc_obj <- performance(Score, "auc")
auc.glm <- Roc_obj@y.values[[1]]
acc_glm <- rep(NA, nsim)
sen_glm <- rep(NA, nsim)
spec_glm <- rep(NA, nsim)
f <- rep(NA, nsim)

for (i in 1:nsim) {
  testID <- sample(n, p, replace = FALSE)
  data.tr <- Cancer[-testID,]
  data.test <- Cancer[testID,]
  Pred2 <- train(diagnosis~., data = data.tr, method = "glm", trControl =
ctrl)
  pred <- predict(Pred2, data.test)
  a <- confusionMatrix(pred, data.test$diagnosis)
  acc_glm[i] <- a$overall[[1]]
  sen_glm[i] <- a$byClass[[1]]
  spec_glm[i] <- a$byClass[[2]]
  f[i] <- a$byClass["F1"]
}

acc.5kcv <- mean(na.omit(acc_glm))
sen.5kcv <- mean(na.omit(sen_glm))
spec.5kcv <- mean(na.omit(spec_glm))
f1 <- mean(na.omit(f))
data.frame(acc = acc.5kcv, sen = sen.5kcv, spec = spec.5kcv, F1 = f1, AUROC =
auc.glm)
```

```
##         acc       sen      spec       F1     AUROC
## 1 0.8140351 0.7842183 0.8321895 0.879723 0.7911382
```

```
summary(pred)
```

```
## B M 
## 2 3
```