Diabetes Risk Prediction (PCA & Stepwise Selection)

MINHCHAU

```
library(dplyr)
library(tidyr)
library(car)
library(psych) # for pairs.panels()
library(factoextra) # for fviz_cluster()
library(ggplot2)
```

Diabetes is a chronic disease characterized by elevated blood glucose levels. A student at the National Institute of Diabetes is studying the diabetes status of a sample of 768 women from a population near Phoenix, Arizona. The dataset aims to provide insights into factors that contribute to the likelihood of having diabetes and can be used can be used to predict diabetes status based on the other variables.

Use the dataset “diabetesdata.csv”. The relevant variables include:

- pregnancies: The number of times the patient has been pregnant.
- glucose: Plasma glucose concentration (in mmol/L) measured during an oral glucose tolerance.
- bloodpressure: Diastolic blood pressure (in mm Hg).
- skinthickness: Thickness of the triceps skin-fold (in mm).
- insulin: Concentration of plasma glucose in (in μU/mL). Elevated insulin levels may indicate insulin resistance.
- BMI: Body Mass Index, a measure of body fat based on weight and height (in kg/m²).
- age : The patient’s age in years.
- diabetesPF: A score assessing the likelihood of diabetes based on family history.
- outcome: The diabetes status, where a value of 0 indicates no diabetes and a value of 1 indicates the presence of diabetes

```
dt = read.csv('diabetesdata.csv')
head(dt)
```

```
##   pregnancies glucose bloodpressure skinthickness insulin  BMI diabetesPF age
## 1           6     148            72            35       0 33.6      0.627  50
## 2           1      85            66            29       0 26.6      0.351  31
## 3           8     183            64             0       0 23.3      0.672  32
## 4           1      89            66            23      94 28.1      0.167  21
## 5           0     137            40            35     168 43.1      2.288  33
## 6           5     116            74             0       0 25.6      0.201  30
##   outcome
## 1       1
## 2       0
## 3       1
## 4       0
## 5       1
## 6       0
```

Section I :

Understand key concepts in PCA for dimensionality reduction.

(3a) Apply Principal Component Analysis (PCA) on the following variables: ‘ `pregnancies'` , `'glucose'` , `'bloodpressure'` , `'skinthickness'` , `'insulin'`,`` `BMI' ``,`'age'` and `'diabetesPF'`.

- Explain what high dimensionality in a dataset means.
- Perform PCA on the selected variables, standardizing them to have a mean of zero and unit variance. Store the resulting PCA object in diabetes.pca. (2 marks)

High dimensionality means your dataset has a lot of features (columns/variables), which can mean dozens, hundreds, or even thousands. - Each new feature adds a new dimension, making the data space more complex. - As dimensions increase, data points become sparser - It becomes harder to visualize, analyze, or model relationships effectively. => We need PCAto reduce dimensionality while keeping the most important information

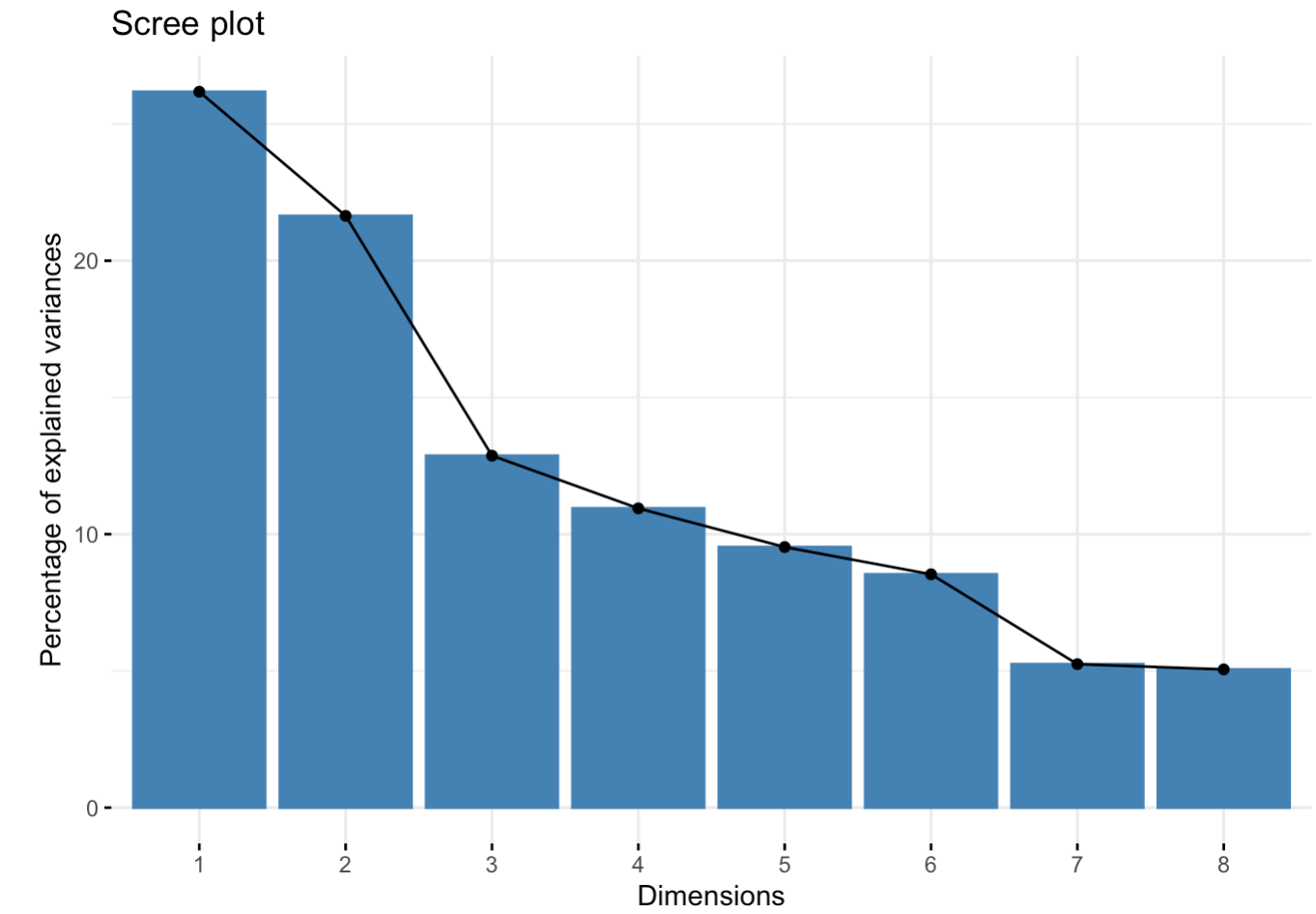
```
pt1 <- dt %>% select(pregnancies, glucose, bloodpressure, skinthickness, insulin, BM
I, age, diabetesPF )
pt2 <- scale(pt1)
pt3 <- prcomp(pt2, center=TRUE, scale.= TRUE)
names(pt3)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
summary(pt3)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.4472 1.3158 1.0147 0.9357 0.87312 0.82621 0.64793
## Proportion of Variance 0.2618 0.2164 0.1287 0.1094 0.09529 0.08533 0.05248
## Cumulative Proportion  0.2618 0.4782 0.6069 0.7163 0.81164 0.89697 0.94944
##                            PC8
## Standard deviation     0.63597
## Proportion of Variance 0.05056
## Cumulative Proportion  1.00000
```

```
fviz_eig(pt3)
```




```
dt22 <- dt
dt22$index <- rownames(dt)
```

(3b) A clinician is concerned that PCA may lead to information loss of the original data.

- Explain the proportion of variance (expressed as a percentage) captured by each Principal Component (PC), specifically from PC1 to PC8.
- Based on the clinician's recommendation to retain at least 60% of the cumulative proportion of variance, determine the minimum number of Principal Components required.

Component Variance (%)

PC1: The 1st principal component explains about 26.18% of the proportion of variance.

PC2: The 2nd principal component explains about 21.64% of the proportion of variance.

PC3: The 3rd principal component explains about 12.87% of the proportion of variance.

PC4: The 4th principal component explains about 10.94% of the proportion of variance.

PC5: The 5th principal component explains about 9.529% of the proportion of variance.

PC6: The 6th principal component explains about 8.533% of the proportion of variance.

PC7: The 7th principal component explains about 5.248% of the proportion of variance.

PC8: The 8th principal component explains about 5.056% of the proportion of variance.

The cumulative proportion of variance > 60 % at PC3. At least 3 Principle Components are needed for the cumulative proportions of variance to be at least 60%.=> minimum number of Principal Components needed is 3

Section II :

Interpret principal components using PC loadings and the biplot.

(3c) Express the first two Principal Components (PC1 and PC2) mathematically as normalized linear combinations of the original variables. (2 marks)

```
rbind(pt3$rotation[, "PC1"], pt3$rotation[, "PC2"])
```

```
##      pregnancies    glucose bloodpressure skinthickness    insulin        BMI
## [1,]  -0.1284321 -0.3930826    -0.3600026    -0.4398243 -0.4350262 -0.4519413
## [2,]   0.5937858  0.1740291     0.1838921    -0.3319653 -0.2507811 -0.1009598
##             age diabetesPF
## [1,] -0.1980271 -0.2706114
## [2,]  0.6205885 -0.1220690
```

```
round(pt3$rotation[, 1:2], 3)
```

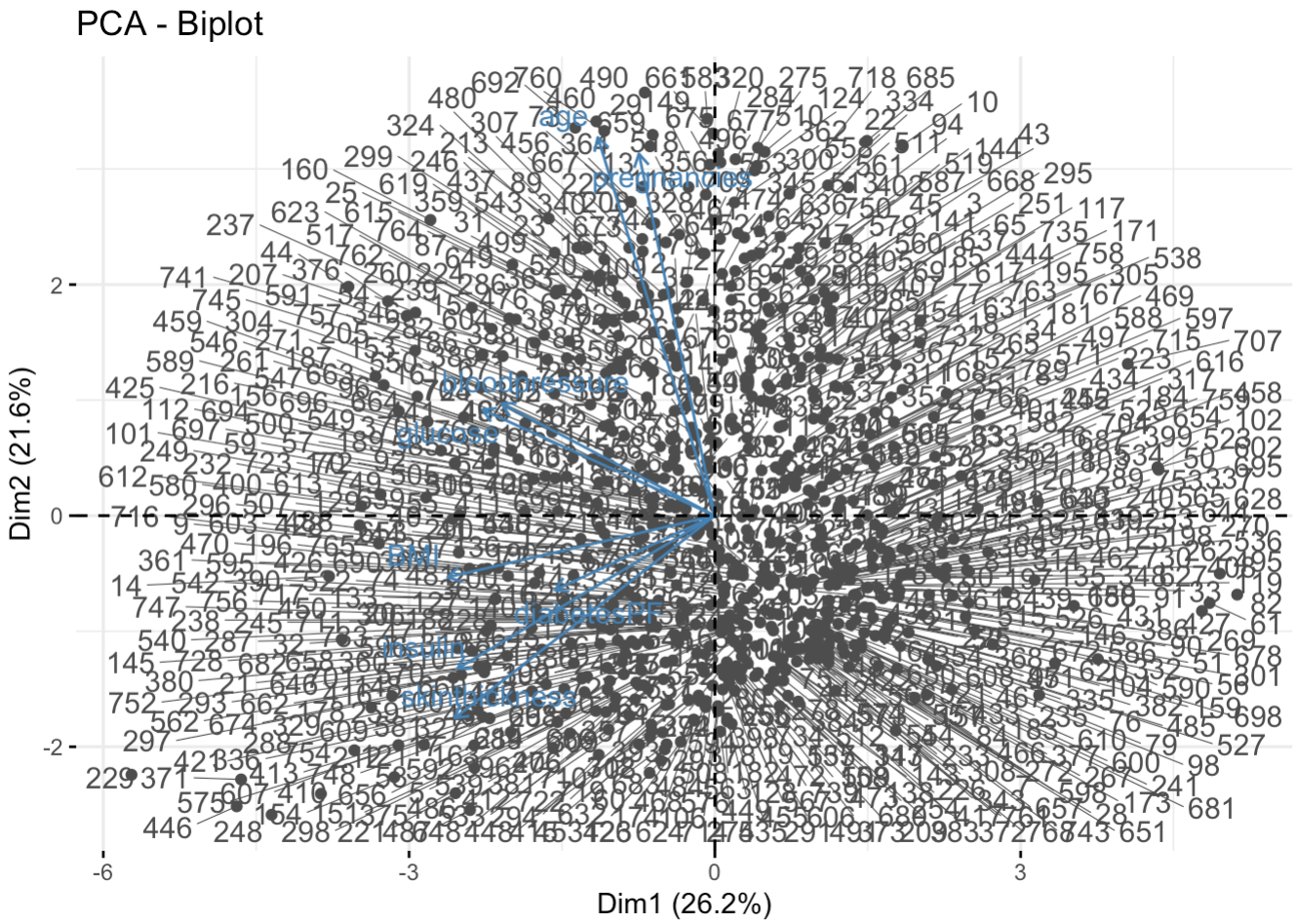
```
##                  PC1    PC2
## pregnancies   -0.128  0.594
## glucose       -0.393  0.174
## bloodpressure -0.360  0.184
## skinthickness -0.440 -0.332
## insulin       -0.435 -0.251
## BMI           -0.452 -0.101
## age           -0.198  0.621
## diabetesPF    -0.271 -0.122
```

PC1 = -0.128.pegnancies -0.393.glucose - 0.360.bloodpressure -0.440.skinthickness -0.435.insulin-0.452.BMI - 0.198.age -0.271.diabetesPF PC2 = 0.594.pegnancies +0.174.glucose + 0.184.bloodpressure -0.332.skinthickness -0.251.insulin - 0.101.BMI + 0.621.age -0.122.diabetesPF

(3d) A student claims that the principal components presented in the biplot lack interpretability. Create a biplot to visualize the PCA results. (2 marks)

- Analyse the principal component loadings and interpret principal component 1 (PC1) in the context of diabetes analysis.
- Using the principal component loadings and biplot, interpret Principal Component 2 (PC2) in the context of diabetes analysis.

```
fviz_pca_biplot(pt3, repel=TRUE, col.var = "steelblue", col.ind="gray30")
```



PC1 The strongest contributors to PC1 are:BMI (-0.452), SkintThickness (-0.440), Insulin (-0.435), Glucose (-0.393) PC1 can be interpreted as a “Risk Factor Component.” People with higher values in glucose, BMI, insulin will have lower PC1 scores. In the biplot, observations with high diabetes risk cluster in the negative PC1 direction. PC1 is useful for identifying individuals at higher risk for Type 2 diabetes (due to physiological indicators).

PC2: can be interpreted as a “Demographic vs. Risk Factor Contrast” Positive loadings: Pregnancies (0.594), Age (0.621), Glucose (0.174) and blood pressure (0.184) Negative loadings: SkintThickness (-0.332), Insulin (-0.251), BMI (-0.101), DiabetesPF (-0.122)

Higher PC2 scores are associated with older individuals with more pregnancies and milder metabolic signs. Lower PC2 scores are linked to younger individuals with higher BMI, more insulin resistance, and higher skin thickness — which are stronger indicators of diabetes risk.

(3e) Fit a logistic regression model with ‘outcome’ as the response variable and the selected principal components from (3b) as predictors.

- Using statistical evidence, explain which principal components are statistically significant. Support your answer with appropriate statistical measures.

Log-odds(Outcome) = -0.76021 - 0.68010 * PC1 + 0.37347 * PC2 + 0.47263 * PC3

Since p-value of PC1, PC2, PC3 components are all smaller than 0.05 (p of PC1 < 2e-16, p of PC2 = 4.27e-09, p of PC3 = 1.92e-07), there is statistically evidence that PC1, PC2, PC3 components are statistically significant to this model at 5% level of significant.

```
logmodel <- glm(dt$outcome ~ pt3$x[,"PC1"] + pt3$x[,"PC2"] + pt3$x[,"PC3"], data = d
t, family = binomial())

summary(logmodel)
```

```
##
## Call:
## glm(formula = dt$outcome ~ pt3$x[, "PC1"] + pt3$x[, "PC2"] +
##     pt3$x[, "PC3"], family = binomial(), data = dt)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.76021    0.08829  -8.610  < 2e-16 ***
## pt3$x[, "PC1"] -0.68010    0.06930  -9.814  < 2e-16 ***
## pt3$x[, "PC2"]  0.37347    0.06359   5.873 4.27e-09 ***
## pt3$x[, "PC3"]  0.47263    0.09076   5.207 1.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 808.76  on 764  degrees of freedom
## AIC: 816.76
##
## Number of Fisher Scoring iterations: 4
```

Section III : Perform model selection using stepwise regression.

(3f) Perform model selection on the logistics regression model using stepwise regression. Use all variables as predictors and ‘outcome’ as the response.

- Fit a full logistic regression model with all predictors and perform backward stepwise selection to remove insignificant predictors. What are the final selected variables?
- Next perform forward stepwise selection, starting with an intercept-only model. What are the final selected variables?

```
backward_model <- glm(outcome ~ ., data= dt, family = binomial)
forward_model <- glm(outcome ~ 1, data =dt, family = binomial)

#Backward stepwise model selection
step(backward_model, direction = 'backward', trace = 1)
```

```
## Start:  AIC=741.45
## outcome ~ pregnancies + glucose + bloodpressure + skinthickness +
##     insulin + BMI + diabetesPF + age
##
##                 Df Deviance    AIC
## - skinthickness  1   723.45 739.45
## - insulin        1   725.19 741.19
## <none>               723.45 741.45
## - age            1   725.97 741.97
## - bloodpressure  1   729.99 745.99
## - diabetesPF     1   733.78 749.78
## - pregnancies    1   738.68 754.68
## - BMI            1   764.22 780.22
## - glucose        1   838.37 854.37
##
## Step:  AIC=739.45
## outcome ~ pregnancies + glucose + bloodpressure + insulin + BMI +
##     diabetesPF + age
##
##                 Df Deviance    AIC
## <none>               723.45 739.45
## - insulin        1   725.46 739.46
## - age            1   725.97 739.97
## - bloodpressure  1   730.13 744.13
## - diabetesPF     1   733.92 747.92
## - pregnancies    1   738.69 752.69
## - BMI            1   768.77 782.77
## - glucose        1   840.87 854.87
```

```
##
## Call:  glm(formula = outcome ~ pregnancies + glucose + bloodpressure +
##     insulin + BMI + diabetesPF + age, family = binomial, data = dt)
##
## Coefficients:
##   (Intercept)    pregnancies        glucose  bloodpressure        insulin
##     -8.405136       0.123172       0.035112      -0.013214      -0.001157
##           BMI     diabetesPF            age
##      0.090089       0.947595       0.014789
##
## Degrees of Freedom: 767 Total (i.e. Null);  760 Residual
## Null Deviance:       993.5
## Residual Deviance: 723.5     AIC: 739.5
```

```
#Forward stepwise model selection
step(forward_model, scope = ~ pregnancies + glucose + bloodpressure + skinthickness +
insulin + BMI + diabetesPF + age,  direction = 'forward', trace = 1)
```

```
## Start:  AIC=995.48
## outcome ~ 1
##
##                 Df Deviance    AIC
## + glucose        1   808.72 812.72
## + BMI            1   920.71 924.71
## + age            1   950.72 954.72
## + pregnancies    1   956.21 960.21
## + diabetesPF     1   970.86 974.86
## + insulin        1   980.81 984.81
## + skinthickness  1   989.19 993.19
## + bloodpressure  1   990.13 994.13
## <none>               993.48 995.48
##
## Step:  AIC=812.72
## outcome ~ glucose
##
##                 Df Deviance    AIC
## + BMI            1   771.40 777.40
## + pregnancies    1   784.95 790.95
## + diabetesPF     1   796.99 802.99
## + age            1   797.36 803.36
## <none>               808.72 812.72
## + skinthickness  1   807.07 813.07
## + insulin        1   807.77 813.77
## + bloodpressure  1   808.59 814.59
##
## Step:  AIC=777.4
## outcome ~ glucose + BMI
##
##                 Df Deviance    AIC
## + pregnancies    1   744.12 752.12
## + age            1   755.68 763.68
## + diabetesPF     1   762.87 770.87
## + insulin        1   767.79 775.79
## + bloodpressure  1   769.07 777.07
## <none>               771.40 777.40
## + skinthickness  1   770.20 778.20
##
## Step:  AIC=752.12
## outcome ~ glucose + BMI + pregnancies
##
##                 Df Deviance    AIC
## + diabetesPF     1   734.31 744.31
## + bloodpressure  1   738.43 748.43
## + age            1   742.10 752.10
## <none>               744.12 752.12
## + insulin        1   742.43 752.43
## + skinthickness  1   743.60 753.60
##
## Step:  AIC=744.31
## outcome ~ glucose + BMI + pregnancies + diabetesPF
##
##                 Df Deviance    AIC
## + bloodpressure  1   728.56 740.56
```

```
## + insulin        1   731.51 743.51
## <none>               734.31 744.31
## + age            1   732.51 744.51
## + skinthickness  1   733.06 745.06
##
## Step:  AIC=740.56
## outcome ~ glucose + BMI + pregnancies + diabetesPF + bloodpressure
##
##                 Df Deviance    AIC
## + age            1   725.46 739.46
## + insulin        1   725.97 739.97
## <none>               728.56 740.56
## + skinthickness  1   728.00 742.00
##
## Step:  AIC=739.46
## outcome ~ glucose + BMI + pregnancies + diabetesPF + bloodpressure +
##     age
##
##                 Df Deviance    AIC
## + insulin        1   723.45 739.45
## <none>               725.46 739.46
## + skinthickness  1   725.19 741.19
##
## Step:  AIC=739.45
## outcome ~ glucose + BMI + pregnancies + diabetesPF + bloodpressure +
##     age + insulin
##
##                 Df Deviance    AIC
## <none>               723.45 739.45
## + skinthickness  1   723.45 741.45
```

```
##
## Call:  glm(formula = outcome ~ glucose + BMI + pregnancies + diabetesPF +
##     bloodpressure + age + insulin, family = binomial, data = dt)
##
## Coefficients:
##   (Intercept)        glucose            BMI    pregnancies     diabetesPF
##     -8.405136       0.035112       0.090089       0.123172       0.947595
## bloodpressure            age        insulin
##     -0.013214       0.014789      -0.001157
##
## Degrees of Freedom: 767 Total (i.e. Null);  760 Residual
## Null Deviance:       993.5
## Residual Deviance: 723.5     AIC: 739.5
```

Backward and Forward Stepwise Selection identified the most significant predictors for the outcome (diabetes). The forward model suggests that the best model is outcome ~ pregnancies + glucose + bloodpressure + insulin + BMI + diabetesPF + age => The forward model excludes skin thickness as a predictor. Based on the backward stepwise model, our best model is outcome ~ glucose + BMI + pregnancies + diabetesPF + bloodpressure + age + insulin. with the Coefficients: glucose BMI pregnancies diabetesPF bloodpressure age
0.035112 0.090089 0.123172 0.947595 -0.013214 0.014789
insulin

-0.001157

=> The model has a reasonable fit (based on the deviance and AIC), and the coefficients tell us the relationship between these predictors and the likelihood of having diabetes.