DSC 424 FINAL REPORT

Diabetes Indicators Dataset

Hai Ha Vu

Non-Technical Summary

Diabetes is an extremely prevalent health issue within the United States that can introduce both a reduced quality of life and financial strain among patients. The Diabetes Health Indicators Dataset is a public domain collection of 253,680 responses to a CDC survey on health-related behaviors and conditions. There are 22 variables within the dataset, one of which indicating whether the person had no diabetes (or diabetes only during pregnancy), prediabetes, or diabetes. The 21 other variables can be used to analyze any trends in the data to identify latent health factors or predict presence or absence of diabetes.

I investigated the use of three different techniques to analyze diabetes with the corresponding health indicators in the dataset. The first two were methods to explore the interrelationships between the health indicators or discover hidden factors within the data. These techniques are called Correspondence Analysis (CA) and Principal Component or Principal Factor Analysis (PCA/PFA). CA can reveal an association between predictor variables and the classes of the response variable. In our case, I attempted to find health indicators most heavily associated with diabetes and prediabetes as opposed to indicators more closely related to people without diabetes. An initial analysis found that conditions such as stroke, heart disease, and difficulty walking are closely related to diabetes in the dataset whereas higher physical activity and having healthcare coverage are associated with no diabetes. Generally, negative health indicators tended to be more associated with the presence of diabetes while positive health indicators tended to be more associated with no diabetes. One surprising exception to this rule was that heavy alcohol consumption was actually most heavily associated with no diabetes rather than presence of diabetes. These results are summarized in **Figure 1** where more heavily associated indicators have smaller radial distance (i.e., smaller angles) to the diabetes class.

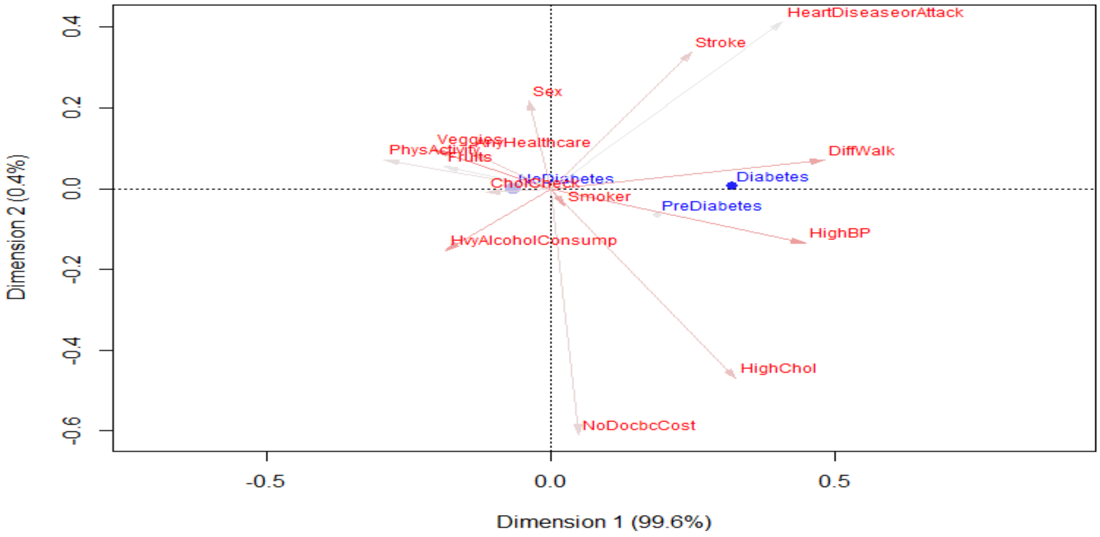


Figure 1: Correspondence analysis radial plot

PFA is a technique that is very closely related to CA but can be used to discover new or latent factors within the dataset as combinations of the existing features. I found two factors from the analysis that seemed to have interpretable meanings. One was comprised of negative health indicators such as high blood pressure or history of stroke, heart disease, or heart attack which may represent prevalent health risks, especially those that are cardiovascular in nature. The other was comprised of opposing negative and positive health indicators which may indicate a factor representing general health with some features contributing to poor health and others contributing to good health.

The third technique that was attempted had a different purpose, that is, for classification and prediction of presence of diabetes. This method is called Linear Discriminant Analysis (LDA) which allows users to train a model for classifying instances and then use it for extrapolated prediction on unknown cases. LDA is another technique that combines variables; however, the combination is used as a classifier to separate datapoints of different classes. Unfortunately, our LDA classifier was not able to perform well at this diabetes classification task and failed to predict even a single prediabetes case. Our hypothesis is that a heavy class imbalance within the diabetes variable partially led to this poor performance. I can visualize this classification task with **Figure 2** which shows the datapoints plotted along axes representing the combined variables that were calculated via LDA. I can see that there is potentially a general difference between diabetes (black points) and no diabetes (green points) but a very large degree of mixing which makes the separation of the two classes quite difficult.

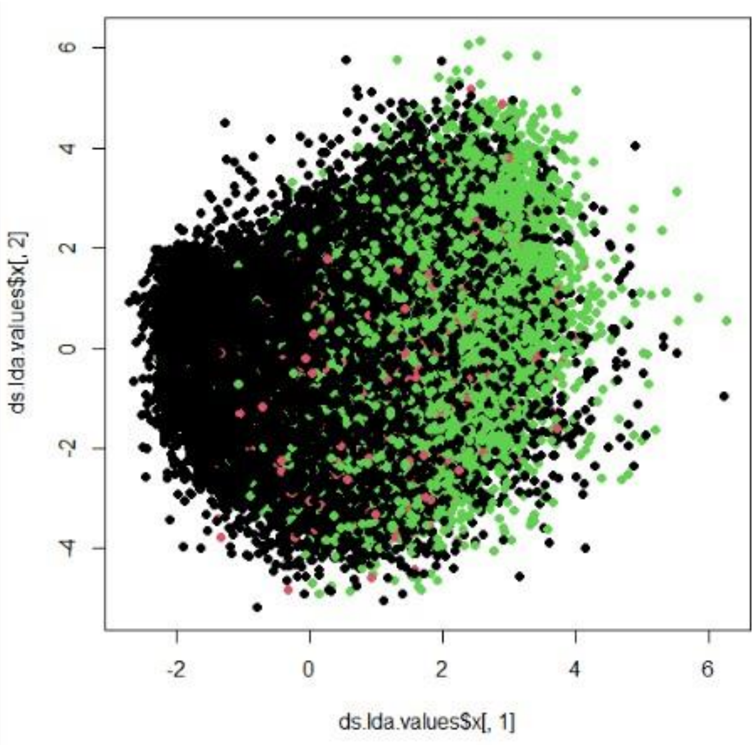


Figure 2: Plot of datapoints along the primary and secondary linear discriminants.

Technical Summary

The diabetes dataset that our group will be analyzing consists of 22 variables, of which 19 (including the response variable) are categorical/ordinal and 3 are continuous. 14 of the categorical variables are binary. Given the nature of the data, I propose two lines exploratory analysis with Principal Component Analysis (PCA) and Correspondence Analysis (CA). Furthermore, I would like to investigatediabetes classification using Linear Discriminant Analysis (LDA). The response variable has 3 categories: 0for no diabetes or diabetes only during pregnancy, 1 for pre-diabetes, and 2 for diabetes.

Principal Factor Analysis

The 22 variables in the diabetes health indicators dataset are all coded as numeric which allows us to use not only correspondence analysis with the categorical variables, but also PCA in order to analyze variable relationships. However, the dataset is comprised of a few different types of variables. 14 are binary, 5 are ordinal, and 3 are continuous. Because of this heterogeneity, several different correlation coefficients needed to be calculated. I used polychoric correlation for ordinal-ordinal pairs, Pearson correlation for continuous- continuous pairs, and polyserial correlation for continuous-ordinal pairs.The response variable was removed andthe remaining features were inputted into R’s hetcor() function to compute all correlations and compile them into a correlation matrix which can be visualized in the correlation plot in **Figure 3**. The correlation plot was ordered by the angle of eigenvectors and reveals potentially 2-3 groupings of the variables. These groupings were kept in mind for the PFA since they may constitute the factors that are calculated.

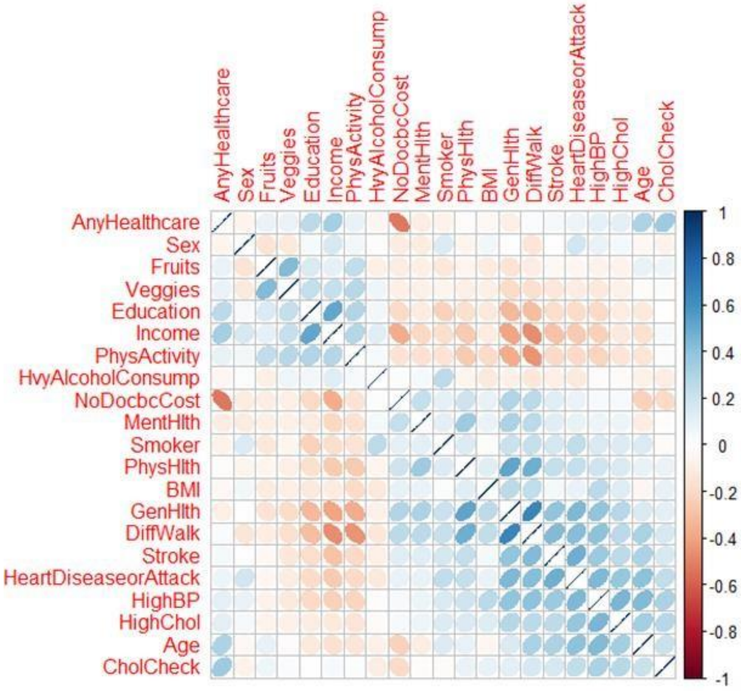


Figure 3: Correlation plot of the 21 independent variables

A correlation test was also conducted to test the significance of these calculated coefficients. Any variable that is highly correlated with a large number of other features may cause issues with the factor analysis because it may be difficult to distinguish its effect from the other features. On the other hand, any variables that are not correlated with the others at all may constitute their own factors. Therefore, both needed to be removed prior to computing the factors. In this case, CholCheck, HvyAlcoholConsump, Sex, and BMI were removed due to having 0 significant correlations with any of the other features.

The new correlation matrix was computed using the remaining 17 variables and used in PCA to determine an appropriate number of factors. The resulting variances of each component were plotted in a scree plot which showed a fairly clear elbow after the second component as seen in **Figure 4a**. Furthermore, a parallel analysis was used to compare variance captured in the diabetes data compared to randomly generated data. This revealed that only the first two factors captured more variance than the noise as seen in **Figure 4b**.

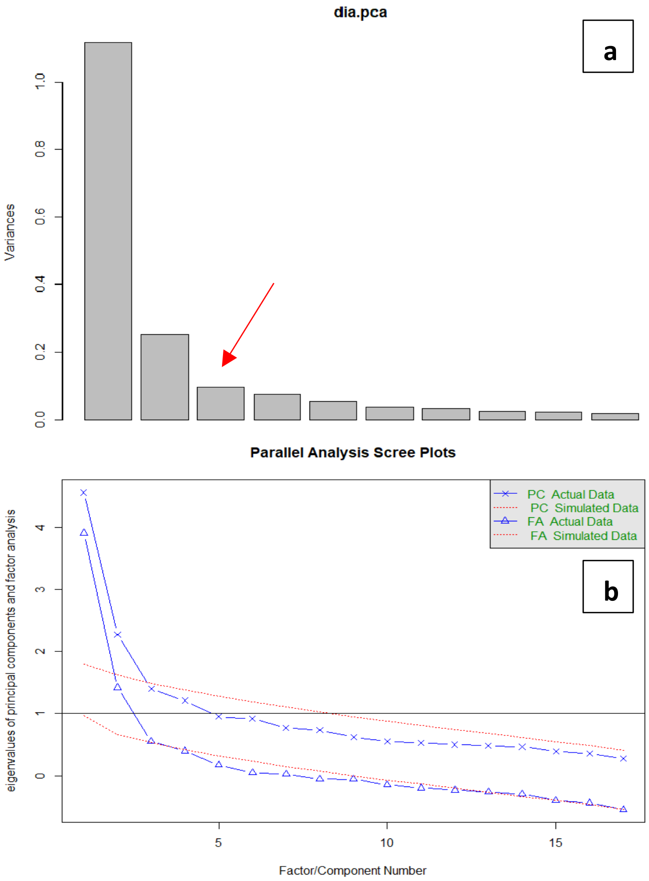


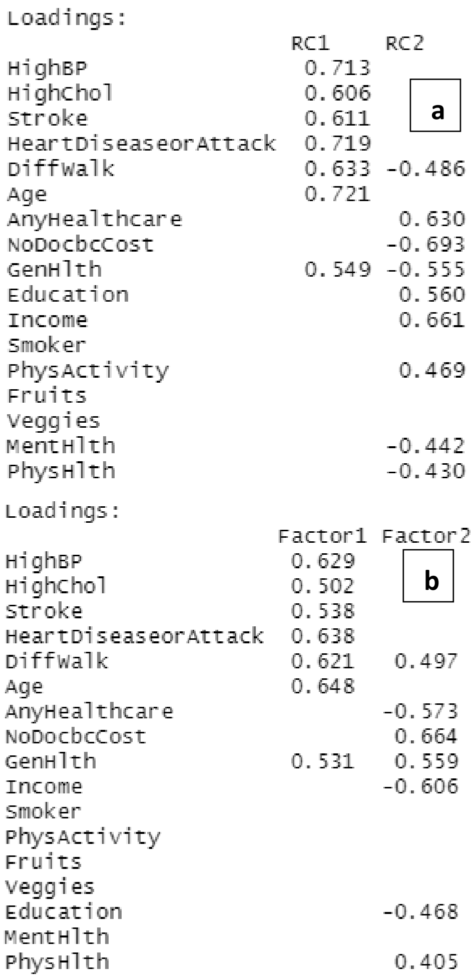
Figure 4: (a) Scree plot by plotting variance of components from prcomp(). (b) Scree plot from parallel analysis of diabetes data compared to noise

These first two components captured 76.13% of the total variance which is sufficient for factor analysis. Thus, two factors were selected when computing principal factors using the R Psych package's principal() function. Varimax rotation was also used in order to improve the interpretability of the PFA. This was more important that preserving the orthogonal components given our goal to discover latent factors rather than to simply reduce dimensionality. This resulted in the loadings that are shown in **Figure 5a**. As a note, all values below 0.4 were filtered out due to their relatively small loading onto the factors.

The result had a high degree of interpretability. Factor 1 had all positive contributions from variables representing poor health indicators. There seemed to be an emphasis on many cardiovascular health-related issues such as high blood pressure, high cholesterol, and stroke. Factor 2 seemed to represent a more general health measure with opposing contributions from good health indicators and poor health indicators. For example, having higher education and income along with healthcare coverage are likely correlated with better preventive or healthcare-seeking behaviors and all had positive contributions. On the other hand, indicators such as difficulty walking and poor mental or physical health had negative contributions. The removed variables can also be added back as their own factors.

Lastly, a confirmatory Common Factor Analysis (CFA) was conducted. PFA is calculated using geometric properties of the data whereas CFA is computed statistically using maximum likelihood techniques. Thus, if similar results can be found using both methods, I can have increased confidence in the factors. Using factanal() in R yielded the factor loadings shown in **Figure 5b.** Factor 1 is strikingly similar between the two methods while Factor 2 has two fewer contributing features from the CFA results. Otherwise, they have similar loading magnitudes and contributions.

As a note, the common factor analysis yielded an extremely high chi-square value with an approximate p-value of 0.0. This means that I must reject the null hypothesis that 2 factors is sufficient. However, when attempting CFA with 10 factors (the maximum allowed for this dataset with factanal), I still receive a p-value of approximately 0.0. I believe that our extremely large dataset is what is contributing to the high chi-squre values being computed.



```
Loadings:
                     RC1    RC2
HighBP               0.713
HighChol             0.606
Stroke               0.611
HeartDiseaseorAttack 0.719
DiffWalk             0.633 -0.486
Age                  0.721
AnyHealthcare               0.630
NoDocbcCost                -0.693
GenHlth              0.549 -0.555
Education                   0.560
Income                      0.661
Smoker
PhysActivity                0.469
Fruits
Veggies
MentHlth                   -0.442
PhysHlth                   -0.430
```

```
Loadings:
                     Factor1 Factor2
HighBP                0.629
HighChol              0.502
Stroke                0.538
HeartDiseaseorAttack  0.638
DiffWalk              0.621   0.497
Age                   0.648
AnyHealthcare                -0.573
NoDocbcCost                   0.664
GenHlth               0.531   0.559
Income                       -0.606
Smoker
PhysActivity
Fruits
Veggies
Education                    -0.468
MentHlth
PhysHlth                      0.405
```

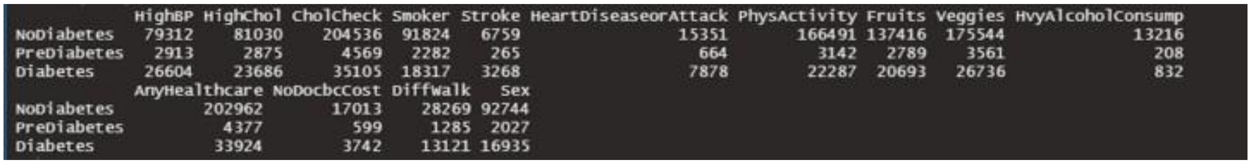
Figure 5: (a) Factor loadings computed using prcomp(). (b) Factor loading computed using factanal(). Both had 2 factors and Varimax rotation.

Correspondence Analysis

Correspondence analysis (CA) and multiple correspondence analysis (MCA) are techniques that are very similar to PCA and factor analysis. Both CA and PCA allow us to summarize patterns in data. CA is used specifically for summarizing and visualizing relative frequencies in tables. The main difference between the two is that CA is well suited for categorical data while PCA is not. MCA differs from CA in that it handles data with more than two dimensions. MCA also excels at dealing with categorical, especially survey data and it is very similar to factor analysis as well. Both are excellent for visualizing multidimensional data into a lower dimensional space.

To begin, I created a frequency table as seen in **Figure 6**. Note that there is a strong frequency imbalance in many of the variables.

I then use the ca() function in R to generate the frequency graph in **Figure 7**. By creating mental scales, I can see that Diabetes has higher frequencies to features such as HighBP and DiffWalk while PreDiabetes has a higher frequency with NoDocbcCost and HighChol. On the other hand, NoDiabetes is more closely associated with PhysAcitivity, Fruits and Veggies, and AnyHealthcare. Generally, more negative health indicators seem to be more closely associated with Diabetes and PreDiabetes while positive health indicators have higher frequencies with NoDiabetes. One glaring exception to this rule is that HvyAlcoholConsump is very highly associated with NoDiabetes. Furthermore, Diabetes and PreDiabetes are fairly close together, especially on Dimension 1 while NoDiabetes is further away and on the other side of the origin.I also note that Dimension 1 captures 99.6% of the variance and Dimension 2 captures .4%.



```
            HighBP HighChol CholCheck Smoker Stroke HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump
NoDiabetes   79312    81030    204536  91824   6759                15351       166491 137416  175544             13216
PreDiabetes   2913     2875      4569   2282    265                  664         3142   2789    3561               208
Diabetes     26604    23686     35105  18317   3268                 7878        22287  20693   26736               832
            AnyHealthcare NoDocbcCost DiffWalk   Sex
NoDiabetes         202962       17013    28269 92744
PreDiabetes          4377         599     1285  2027
Diabetes            33924        3742    13121 16935
```

Figure 6: Frequency table of diabetes classes by binary indicators in the dataset

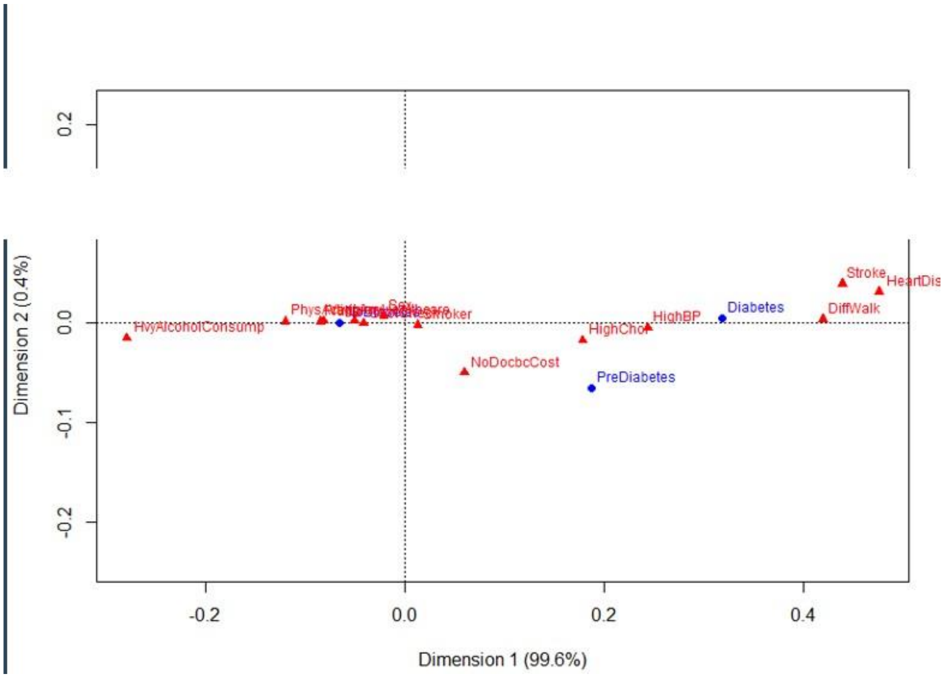
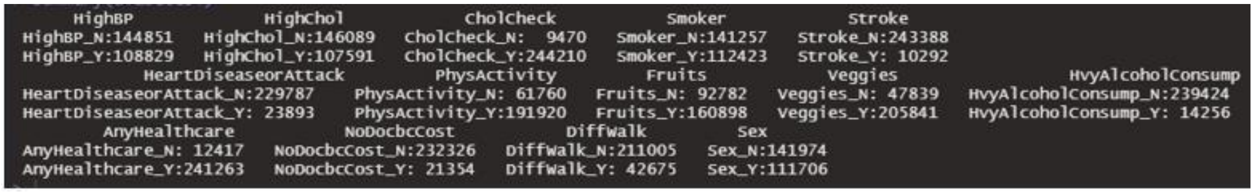


Figure 7: Correspondence Analysis plot using frequency table from Figure 6

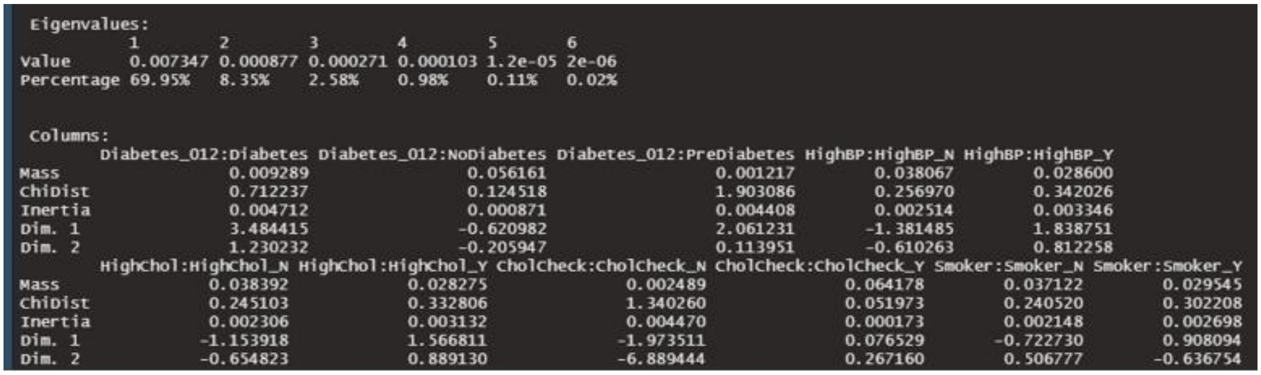
I then moved on to attempt MCA. The preprocessing/transformation involved converting all of the 14 binary categorical columns into descriptive factor columns. The resulting frequency table in **Figure 8** is slightly different than the original frequency table.



```
      HighBP            HighChol            CholCheck           Smoker             Stroke
 HighBP_N:144851   HighChol_N:146089   CholCheck_N:  9470   Smoker_N:141257   Stroke_N:243388
 HighBP_Y:108829   HighChol_Y:107591   CholCheck_Y:244210   Smoker_Y:112423   Stroke_Y: 10292
           HeartDiseaseorAttack           PhysActivity          Fruits           Veggies              HvyAlcoholConsump
 HeartDiseaseorAttack_N:229787    PhysActivity_N: 61760   Fruits_N: 92782   Veggies_N: 47839   HvyAlcoholConsump_N:239424
 HeartDiseaseorAttack_Y: 23893    PhysActivity_Y:191920   Fruits_Y:160898   Veggies_Y:205841   HvyAlcoholConsump_Y: 14256
        AnyHealthcare           NoDocbcCost          DiffWalk          Sex
 AnyHealthcare_N: 12417   NoDocbcCost_N:232326   DiffWalk_N:211005   Sex_N:141974
 AnyHealthcare_Y:241263   NoDocbcCost_Y: 21354   DiffWalk_Y: 42675   Sex_Y:111706
```

Figure 8: Frequency table for multiple correspondence analysis

I then created a summary table in order to see the components generated and the variance capturedas seen in **Figure 9**.



```
 Eigenvalues:
            1        2        3        4        5       6
Value      0.007347 0.000877 0.000271 0.000103 1.2e-05 2e-06
Percentage 69.95%   8.35%    2.58%    0.98%    0.11%   0.02%


 Columns:
        Diabetes_012:Diabetes Diabetes_012:NoDiabetes Diabetes_012:PreDiabetes HighBP:HighBP_N HighBP:HighBP_Y
Mass                 0.009289                0.056161                 0.001217        0.038067        0.028600
ChiDist              0.712237                0.124518                 1.903086        0.256970        0.342026
Inertia              0.004712                0.000871                 0.004408        0.002514        0.003346
Dim. 1               3.484415               -0.620982                 2.061231       -1.381485        1.838751
Dim. 2               1.230232               -0.205947                 0.113951       -0.610263        0.812258
        HighChol:HighChol_N HighChol:HighChol_Y CholCheck:CholCheck_N CholCheck:CholCheck_Y Smoker:Smoker_N Smoker:Smoker_Y
Mass               0.038392            0.028275              0.002489              0.064178        0.037122        0.029545
ChiDist            0.245103            0.332806              1.340260              0.051973        0.240520        0.302208
Inertia            0.002306            0.003132              0.004470              0.000173        0.002148        0.002698
Dim. 1            -1.153918            1.566811             -1.973511              0.076529       -0.722730        0.908094
Dim. 2            -0.654823            0.889130             -6.889444              0.267160        0.506777       -0.636754
```

Figure 9: MCA summary table

Another frequency graph was created in **Figure 10**. This graph is slightly different than the previous CA graph in that it shows Diabetes having higher frequencies with heart disease and stroke.
Both techniques were helpful in visualizing the variables and their frequencies. The factors generated also lined up with the factors generated in the previous PFA section of our project. MCA was able to capture 78% of the variance in the first 2 components. Based on the summary, 2 or 3 components seem appropriate which aligns with the PCA analysis.

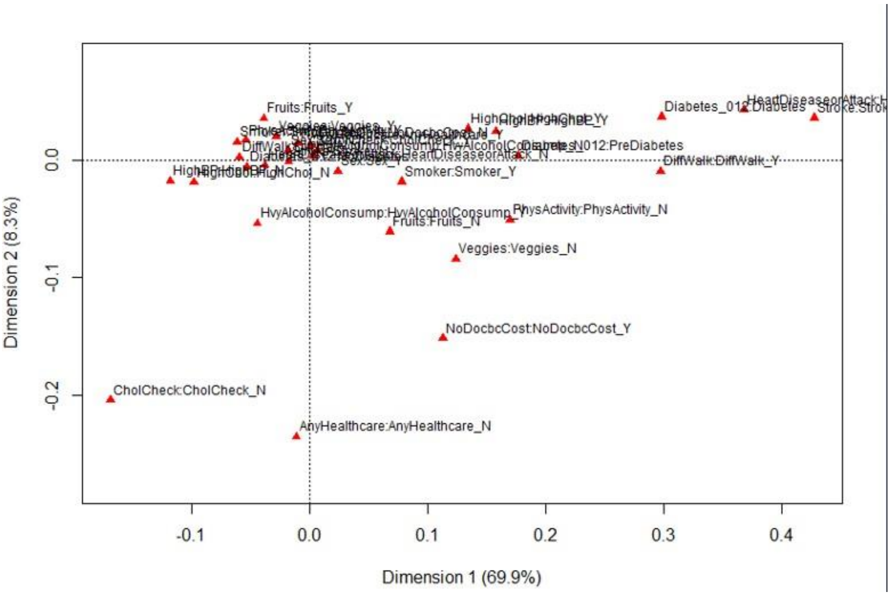
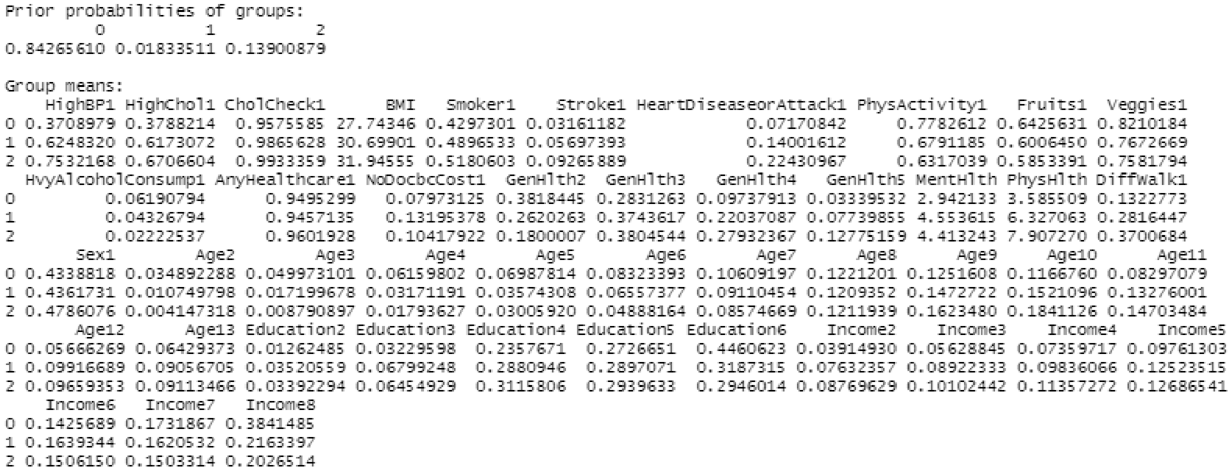


Figure 10: MCA frequency plot

Linear Discriminant Analysis

Firstly, I code all categorical variables as binary using dummy variables and then split the data set via holdout into 80% training and 20% testing. LDA is then computed on all indicators for the training set as seen in **Figure 11**. Looking at the output, I get a set of prior probabilities which is the samples that fall into each category. There is a large class imbalance with the majority of cases falling into type 0 (84%), very little of type 1 (1%), and a moderate amount of type 2 (13%). The group means are the means in each class of the independent variables. This is what LDA is attempting to separate.



```
Prior probabilities of groups:
         0          1          2 
0.84265610 0.01833511 0.13900879 

Group means:
    HighBP1 HighChol1 CholCheck1      BMI   Smoker1    Stroke1 HeartDiseaseorAttack1 PhysActivity1   Fruits1  Veggies1
0 0.3708979 0.3788214  0.9575585 27.74346 0.4297301 0.03161182            0.07170842     0.7782612 0.6425631 0.8210184
1 0.6248320 0.6173072  0.9865628 30.69901 0.4896533 0.05697393            0.14001612     0.6791185 0.6006450 0.7672669
2 0.7532168 0.6706604  0.9933359 31.94555 0.5180603 0.09265889            0.22430967     0.6317039 0.5853391 0.7581794
  HvyAlcoholConsump1 AnyHealthcare1 NoDocbcCost1  GenHlth2  GenHlth3   GenHlth4   GenHlth5 MentHlth PhysHlth DiffWalk1
0         0.06190794      0.9495299   0.07973125 0.3818445 0.2831263 0.09737913 0.03339532 2.942133 3.585509 0.1322773
1         0.04326794      0.9457135   0.13195378 0.2620263 0.3743617 0.22037087 0.07739855 4.553615 6.327063 0.2816447
2         0.02222537      0.9601928   0.10417922 0.1800007 0.3804544 0.27932367 0.12775159 4.413243 7.907270 0.3700684
       Sex1        Age2        Age3       Age4       Age5       Age6       Age7      Age8      Age9     Age10      Age11
0 0.4338818 0.034892288 0.049973101 0.06159802 0.06987814 0.08323393 0.10609197 0.1221201 0.1251608 0.1166760 0.08297079
1 0.4361731 0.010749798 0.017199678 0.03171191 0.03574308 0.06557377 0.09110454 0.1209352 0.1472722 0.1521096 0.13276001
2 0.4786076 0.004147318 0.008790897 0.01793627 0.03005920 0.04888164 0.08574669 0.1211939 0.1623480 0.1841126 0.14703484
       Age12      Age13 Education2 Education3 Education4 Education5 Education6    Income2    Income3    Income4    Income5
0 0.05666269 0.06429373 0.01262485 0.03229598  0.2357671  0.2726651  0.4460623 0.03914930 0.05628845 0.07359717 0.09761303
1 0.09916689 0.09056705 0.03520559 0.06799248  0.2880946  0.2897071  0.3187315 0.07632357 0.08922333 0.09836066 0.12523515
2 0.09659353 0.09113466 0.03392294 0.06454929  0.3115806  0.2939633  0.2946014 0.08769629 0.10102442 0.11357272 0.12686541
    Income6   Income7   Income8
0 0.1425689 0.1731867 0.3841485
1 0.1639344 0.1620532 0.2163397
2 0.1506150 0.1503314 0.2026514
```

Figure 11: LDA summary from initial attempt

Separation of three classes can be accomplished using 2 components. The first discriminant captures 99.14% of the trace while the second captures 0.86%. I then take the LDA scores and plot them in a histogram across the three diabetes classes. The result is displayed in **Figure 12** where I can see that there is very little separation between the three diabetes groups across both linear discriminants.

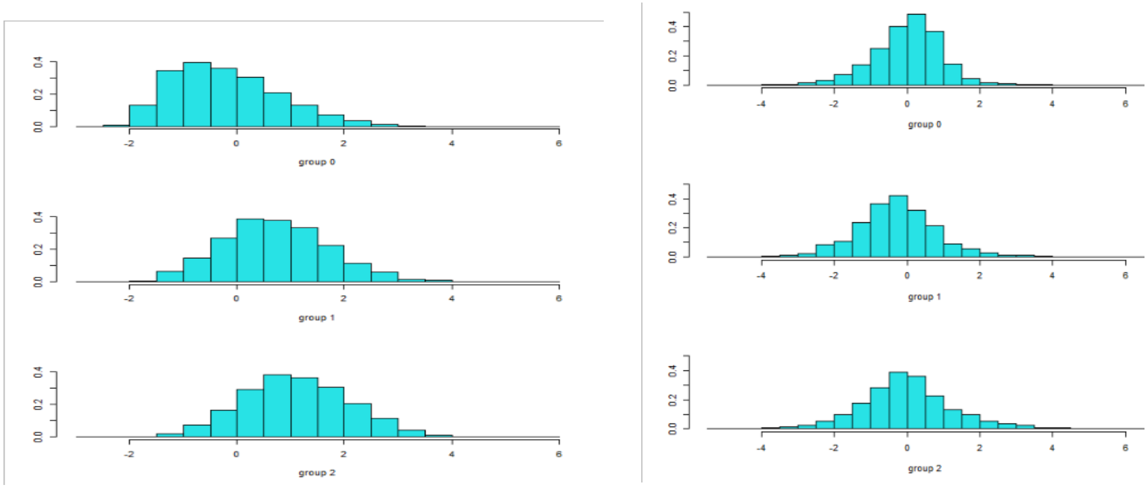


Figure 12: Histogram of scores across linear discriminant 1 (left) and 2 (right)

Another way to visualize this is to plot the scores along the two linear discriminants which can be seen in **Figure 13**. While there are perhaps general groupings of points between Diabetes (black) and No Diabetes (green), there is a large degree of mixing. Furthermore, pre-diabetes data is quite homogenously mixed among the other two classes.

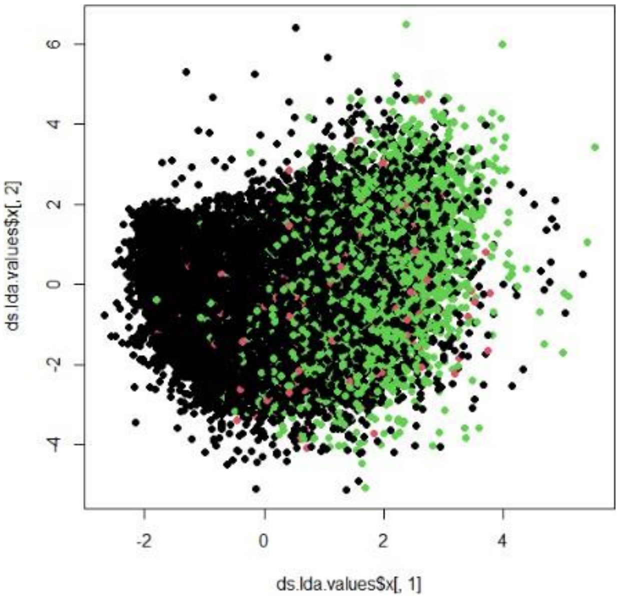


Figure 13: Scores from LD1 and LD2 plotted against each other

When I calculate the confusion matrix on the test set (see **Figure 14**), I see relatively poor performance in classifying all three classes.

	Predicted Class		
Actual Class	0	1	2
0	41022	0	1669
1	788	0	122
2	5389	0	1746

Figure 14: Confusion matrix on LDA test set

Class 0 appears to have the strongest performance of the three while class 2 has poor performance and the model failing to make a single class 1 prediction altogether. I suspect that the heavy class imbalance is contributing to this behavior since the vast majority of cases are class 0 while class 1 is the heavy minority.

Because pre-diabetes is not well separated from either class, future work could involve removing that class entirely from the dataset and focusing on the two class problem instead. Furthermore, I can investigate over- or undersampling for class balance along with data augmentation techniques such as SMOTE for creating synthetic data.

Appendix: R code

PFA

```
# Importing and Preparing Data
diabetes = read.csv('diabetes_health_indicators.csv')
diabetes_cat = diabetes[, c(1:4, 6:15, 18:22)]
diabetes_cat[] = lapply(diabetes_cat, factor)
diabetes_num = diabetes[, c(5, 16:17)]
diabetes = cbind(diabetes_cat, diabetes_num)
diabetes2 = diabetes[-1]
summary(diabetes2)

# Libraries
library(corrplot)
library(MASS)
library(psych)
library(polycor)

# Correlation analysis
cat.cor = hetcor(diabetes2)
print(cat.cor)
corrplot(cat.cor$correlations, method='ellipse', order='AOE')
catcortest = corr.test(cat.cor$correlations, adjust='none')
catcortest.clean = ifelse(catcortest$p < 0.01, T, F)
colSums(catcortest.clean) - 1

# Removing variables with no significant correlations and re-computing correlation matrix
diabetes3 = subset(diabetes2, select=-c(CholCheck, BMI, Sex, HvyAlcoholConsump))
cat.cor2 = hetcor(diabetes3)

# PCA to determine number of factors
dia.pca = prcomp(cat.cor2$correlations)
print(dia.pca)
summary(dia.pca)
plot(dia.pca)
pa = fa.parallel(cat.cor2$correlations, n.iter=500)

# PFA with principal()
dia.prin = principal(cat.cor2$correlations, nfactors=2)
print(dia.prin$loadings, cutoff=0.4, sort=T)

# CFA for confirmatory analysis + chi-square
dia.fact = factanal(cov=cat.cor2$correlations, factors=2, n.obs=253680)
print(dia.fact$loadings, cutoff=0.4, sort=T)
print(dia.fact)
```

LDA

```
library(tidyverse)
library(corrplot)
library(plyr)
library(ggplot2)
library(RCurl)
library(psych)
library(MASS)
library(caret)
library(car)

# Load and explore our data
ds = read.csv("diabetes_012_health_indicators_BRFSS2015.csv")
head(ds)

# Transform some variables to factors
ds$Diabetes_012 = as.factor(ds$Diabetes_012)
ds$GenHlth = as.factor(ds$GenHlth)
ds$Age = as.factor(ds$Age)
ds$Education = as.factor(ds$Education)
ds$Income = as.factor(ds$Income)

ds$HighBP = as.factor(ds$HighBP)
ds$HighChol = as.factor(ds$HighChol)
ds$CholCheck = as.factor(ds$CholCheck)
ds$Smoker = as.factor(ds$Smoker)
ds$Stroke = as.factor(ds$Stroke)
ds$HeartDiseaseorAttack = as.factor(ds$HeartDiseaseorAttack)
ds$PhysActivity = as.factor(ds$PhysActivity)
ds$Fruits = as.factor(ds$Fruits)
ds$Veggies = as.factor(ds$Veggies)
ds$HvyAlcoholConsump = as.factor(ds$HvyAlcoholConsump)
ds$AnyHealthcare = as.factor(ds$AnyHealthcare)
ds$NoDocbcCost = as.factor(ds$NoDocbcCost)
ds$DiffWalk = as.factor(ds$DiffWalk)
ds$Sex = as.factor(ds$Sex)

summary(ds$Diabetes_012)  #three types 0-no diabetes or only during pregnancy
                          #1-prediabetes and 2-diabetes

# Try an initial lda on everything
ds.lda = lda(Diabetes_012 ~ ., data=ds)
# Look at the output
print(ds.lda)
# Print the scaling:
print(ds.lda$scaling[order(ds.lda$scaling[, 1]), ])
print(ds.lda$scaling[order(ds.lda$scaling[, 2]), ])
# Look at the separation
```

```
ds.lda.values = predict(ds.lda)
ldahist(data=ds.lda.values$x[, 1], g=ds$Diabetes_012)
ldahist(data=ds.lda.values$x[, 2], g=ds$Diabetes_012)
# Plot the transformed data:
plot(ds.lda.values$x[, 1], ds.lda.values$x[, 2], col=ds$Diabetes_012, pch=16)
# Compute a confusion matrix
table(ds$Diabetes_012, ds.lda.values$class)
confusionMatrix(ds$Diabetes_012, ds.lda.values$class)

# Now, let's separate into test and training set:
s = sample(nrow(ds), nrow(ds) * .8)
dsTrain = ds[s, ]
dsTest = ds[-s, ]

# Build the model on the training set
ds.lda.train = lda(Diabetes_012 ~ ., data=dsTrain)
# Look at the output
ds.lda.train
# Print the scaling:
print(ds.lda.train$scaling[order(ds.lda.train$scaling[, 1]), ])
# Predict on the test set
ds.lda.values = predict(ds.lda.train, dsTest)
# Look at histograms:
ldahist(data=ds.lda.values$x[, 1], g=dsTest$Diabetes_012)
ldahist(data=ds.lda.values$x[, 2], g=dsTest$Diabetes_012)
# Plot the transformed data so I can see the classification
plot(ds.lda.values$x[, 1], ds.lda.values$x[, 2], col=dsTest$Diabetes_012, pch=16)
# Compute a confusion matrix
table(dsTest$Diabetes_012, ds.lda.values$class)
confusionMatrix(dsTest$Diabetes_012, ds.lda.values$class)
```

CA/MCA

```
library("FactoMineR")
library("factoextra")
library(ca)
library(dplyr)

################## Pre Processing
diabetes =
read.csv("C:/Users/devjrr/Documents/DSC424/FinalProject/diabetes_012_health_indicators/diabetes_0
12_health_indicators_BRFSS2015.csv")
head(diabetes)

########## Convert binary to factors
```

```
diabetes4 = subset(diabetes, select =
c(Diabetes_012,HighBP,HighChol,CholCheck,Smoker,Stroke,HeartDiseaseorAttack,PhysActivity,Fruits,Veggies,HvyAlcoholConsump,AnyHealthcare,NoDocbcCost,DiffWalk,Sex) )

diabetes4$Diabetes_012[diabetes4$Diabetes_012 == 0] <- "NoDiabetes"
diabetes4$Diabetes_012[diabetes4$Diabetes_012 == 1] <- "PreDiabetes"
diabetes4$Diabetes_012[diabetes4$Diabetes_012 == 2] <- "Diabetes"
diabetes4$Diabetes_012 <- as.factor(diabetes4$Diabetes_012)
diabetes4$HighBP[diabetes4$HighBP == 1] <- "HighBP_Y"
diabetes4$HighBP[diabetes4$HighBP == 0] <- "HighBP_N"
diabetes4$HighBP <- as.factor(diabetes4$HighBP)
diabetes4$HighChol[diabetes4$HighChol == 1] <- "HighChol_Y"
diabetes4$HighChol[diabetes4$HighChol == 0] <- "HighChol_N"
diabetes4$HighChol <- as.factor(diabetes4$HighChol)
diabetes4$CholCheck[diabetes4$CholCheck == 1] <- "CholCheck_Y"
diabetes4$CholCheck[diabetes4$CholCheck == 0] <- "CholCheck_N"
diabetes4$CholCheck <- as.factor(diabetes4$CholCheck)
diabetes4$Smoker[diabetes4$Smoker == 1] <- "Smoker_Y"
diabetes4$Smoker[diabetes4$Smoker== 0] <- "Smoker_N"
diabetes4$Smoker <- as.factor(diabetes4$Smoker)
diabetes4$Stroke[diabetes4$Stroke == 1] <- "Stroke_Y"
diabetes4$Stroke[diabetes4$Stroke == 0] <- "Stroke_N"
diabetes4$Stroke <- as.factor(diabetes4$Stroke)
diabetes4$HeartDiseaseorAttack[diabetes4$HeartDiseaseorAttack == 1] <- "HeartDiseaseorAttack_Y"
diabetes4$HeartDiseaseorAttack[diabetes4$HeartDiseaseorAttack == 0] <- "HeartDiseaseorAttack_N"
diabetes4$HeartDiseaseorAttack <- as.factor(diabetes4$HeartDiseaseorAttack)
diabetes4$PhysActivity[diabetes4$PhysActivity == 1] <- "PhysActivity_Y"
diabetes4$PhysActivity[diabetes4$PhysActivity == 0] <- "PhysActivity_N"
diabetes4$PhysActivity <- as.factor(diabetes4$PhysActivity)
diabetes4$Fruits[diabetes4$Fruits == 1] <- "Fruits_Y"
diabetes4$Fruits[diabetes4$Fruits == 0] <- "Fruits_N"
diabetes4$Fruits <- as.factor(diabetes4$Fruits)
diabetes4$Veggies[diabetes4$Veggies == 1] <- "Veggies_Y"
diabetes4$Veggies[diabetes4$Veggies == 0] <- "Veggies_N"
diabetes4$Veggies <- as.factor(diabetes4$Veggies)
diabetes4$HvyAlcoholConsump[diabetes4$HvyAlcoholConsump == 1] <- "HvyAlcoholConsump_Y"
diabetes4$HvyAlcoholConsump[diabetes4$HvyAlcoholConsump == 0] <- "HvyAlcoholConsump_N"
diabetes4$HvyAlcoholConsump <- as.factor(diabetes4$HvyAlcoholConsump)
diabetes4$AnyHealthcare[diabetes4$AnyHealthcare == 1] <- "AnyHealthcare_Y"
diabetes4$AnyHealthcare[diabetes4$AnyHealthcare == 0] <- "AnyHealthcare_N"
diabetes4$AnyHealthcare <- as.factor(diabetes4$AnyHealthcare)
diabetes4$NoDocbcCost[diabetes4$NoDocbcCost == 1] <- "NoDocbcCost_Y"
diabetes4$NoDocbcCost[diabetes4$NoDocbcCost == 0] <- "NoDocbcCost_N"
diabetes4$NoDocbcCost <- as.factor(diabetes4$NoDocbcCost)
diabetes4$DiffWalk[diabetes4$DiffWalk == 1] <- "DiffWalk_Y"
diabetes4$DiffWalk[diabetes4$DiffWalk == 0] <- "DiffWalk_N"
diabetes4$DiffWalk <- as.factor(diabetes4$DiffWalk)
diabetes4$Sex[diabetes4$Sex == 1] <- "Sex_Y"
```

```
diabetes4$Sex[diabetes4$Sex == 0] <- "Sex_N"
diabetes4$Sex <- as.factor(diabetes4$Sex)

################################################### CA

diabetes8 = subset(diabetes, select =
c(Diabetes_012,HighBP,HighChol,CholCheck,Smoker,Stroke,HeartDiseaseorAttack,PhysActivity,Fruits,Veggies,HvyAlcoholConsump,AnyHealthcare,NoDocbcCost,DiffWalk,Sex) )
finald = diabetes8 %>%
  group_by(Diabetes_012) %>%
  summarise(across(everything(), sum), .groups = 'drop')
cadf = as.data.frame(finald)
head(cadf)
cadf$Diabetes_012 = NULL
rownames(cadf) = c( "NoDiabetes", "PreDiabetes","Diabetes")
head(cadf)
fit = ca(cadf)
fit
plot(fit)
plot(fit, mass=T, contrib="absolute",
     map="rowgreen", arrows=c(F, T))

###################### Multiple CA
summary(diabetes4)
dmca <- mjca(diabetes4)
print(dmca)
plot(dmca)
```