Linear Regression

Diabetes Diagnose

radmehr karimian

July 4, 2023

1. Although there are so many medical reasons behind choosing these features we just plot some of the features mean for visualize that diabetes have a relationship with these parameters:

For Example Consider these:

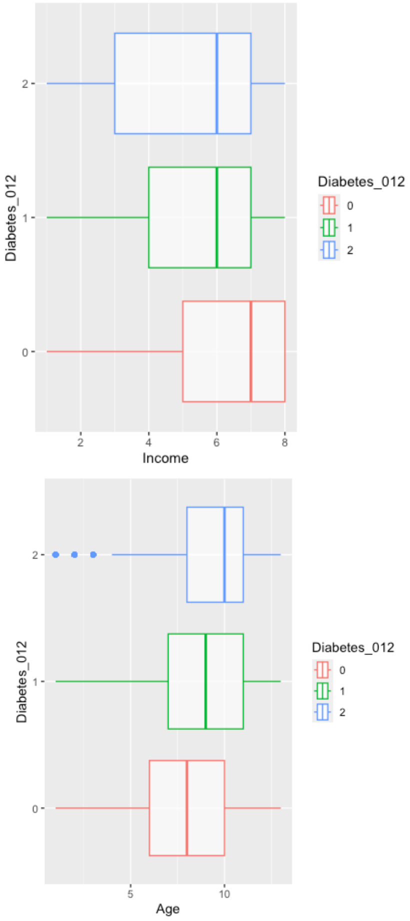


Fig1:Mean of each type of diabetes for some features

For dummy variables, it’s not a good idea to visualize their effect like this, so we have another idea which is showing the percentage of each type of diabetes for each state.

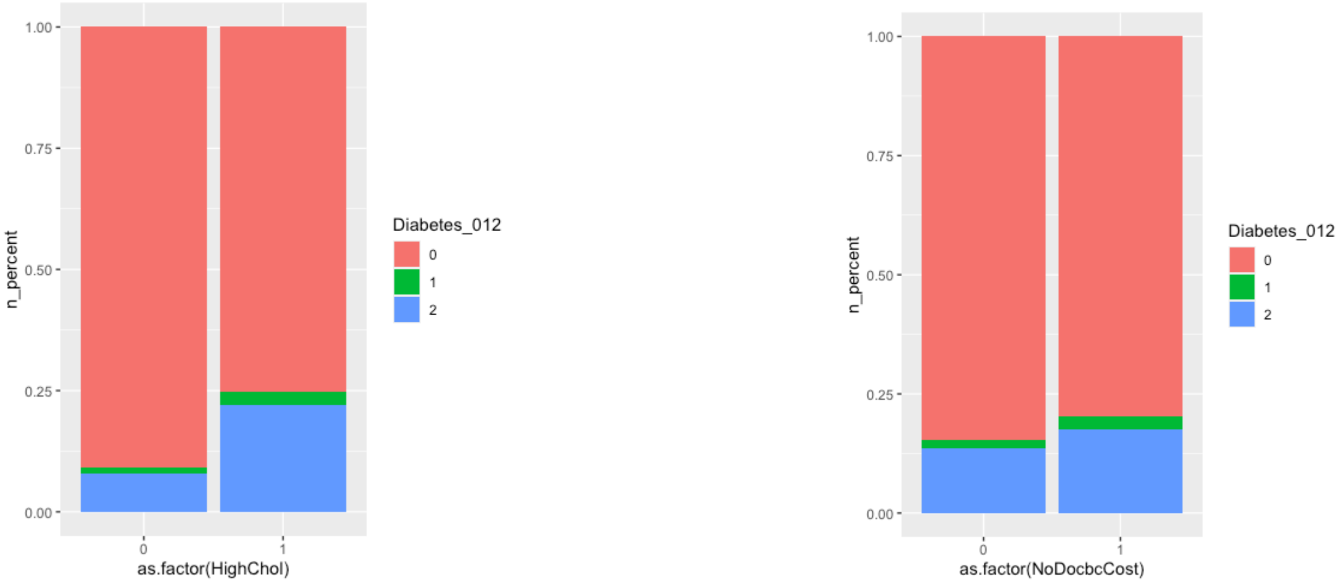


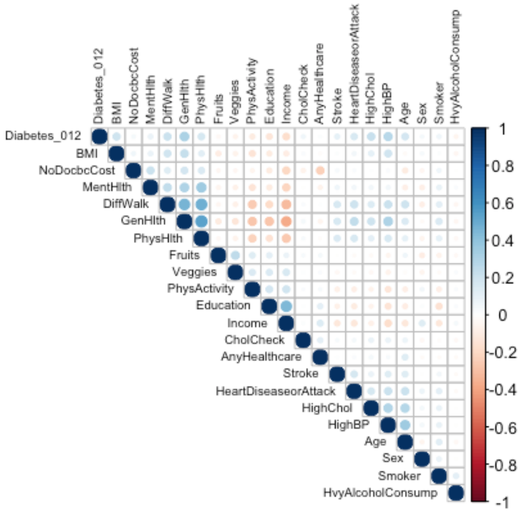
Fig2: percentage of each type of diabetes for dummy variables

So overall, we can say that yes there is a relationship between our features and diabetes!

2-3.

We want to answer these questions together.

First, we want to check the covariance matrix:

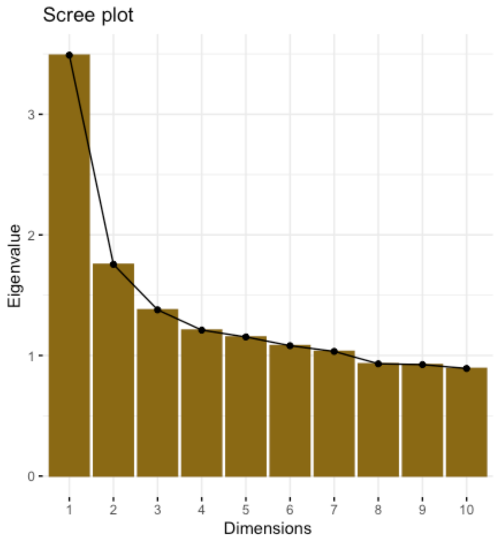


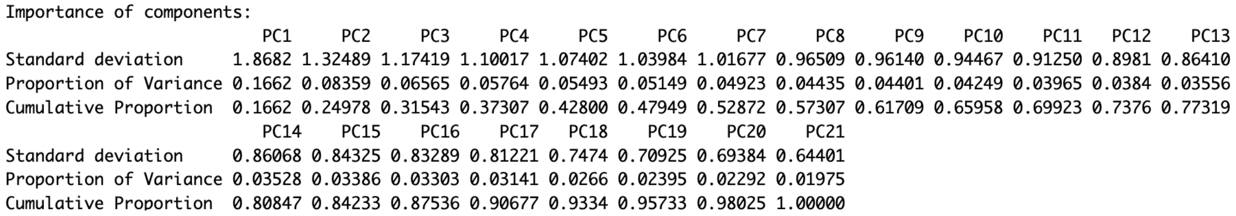
As you can see, there are some serious correlations between some features.

We should consider this when we want to use feature selection!

On the other hand, we want to use PCA for feature selection,

There are some many different approach for that, like:





```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11   PC12    PC13
Standard deviation     1.8682 1.32489 1.17419 1.10017 1.07402 1.03984 1.01677 0.96509 0.96140 0.94467 0.91250 0.8981 0.86410
Proportion of Variance 0.1662 0.08359 0.06565 0.05764 0.05493 0.05149 0.04923 0.04435 0.04401 0.04249 0.03965 0.0384 0.03556
Cumulative Proportion  0.1662 0.24978 0.31543 0.37307 0.42800 0.47949 0.52872 0.57307 0.61709 0.65958 0.69923 0.7376 0.77319
                          PC14    PC15    PC16    PC17   PC18    PC19    PC20    PC21
Standard deviation     0.86068 0.84325 0.83289 0.81221 0.7474 0.70925 0.69384 0.64401
Proportion of Variance 0.03528 0.03386 0.03303 0.03141 0.0266 0.02395 0.02292 0.01975
Cumulative Proportion  0.80847 0.84233 0.87536 0.90677 0.9334 0.95733 0.98025 1.00000
```

Fig3 : scree plot

A scree plot is a graphical method used to determine the number of principal components or factors to retain in a data analysis. It is created by plotting the eigenvalues of each component or factor in descending order against their component or factor number. The plot usually shows a steep drop in eigenvalues for the first few components or factors, followed by a leveling off or gradual decrease in eigenvalues for the remaining components or factors.

The scree plot is helpful in determining the number of components or factors to retain because it shows the point at which the eigenvalues start to level off, indicating that the additional components or factors explain very little additional variance in the data. The number of components or factors to retain is typically chosen at the point where the eigenvalues start to level off, or at the point of the "elbow" in the scree plot.

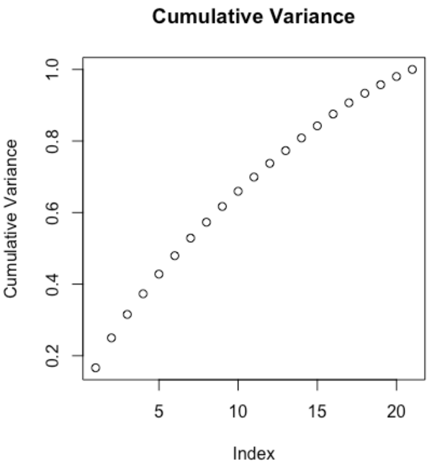
For this plot scree plot suggests 7 for number of PCA.

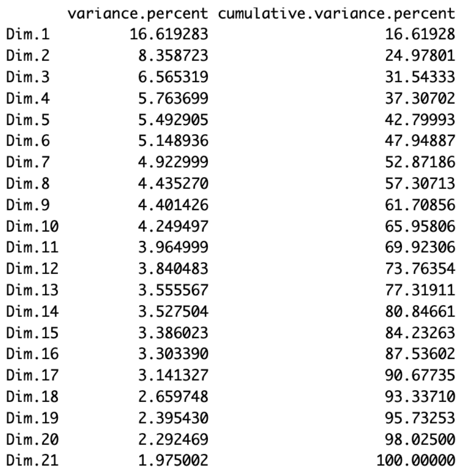
Next we have:

Cumulative variance is a measure of the amount of variance in a dataset that is explained by a set of variables or components. In data analysis, it is often used in principal component analysis (PCA) or factor analysis to determine the number of principal components or factors to retain for modeling.

The cumulative variance is calculated by adding up the individual variances explained by each component or factor in a dataset, in order of importance. The cumulative variance is often plotted as a graph, with the number of components or factors on the x-axis and the cumulative variance on the y-axis. The graph typically shows a steep increase in cumulative variance for the first few components or factors, followed by a leveling off as the additional components or factors explain less and less variance.

The cumulative variance plot is useful in determining the number of components or factors to retain, as it shows the proportion of variance in the data that is explained by the retained components or factors. The goal is often to retain enough components or factors to explain a high percentage of the total variance in the data, while avoiding overfitting by retaining too many components or factors.

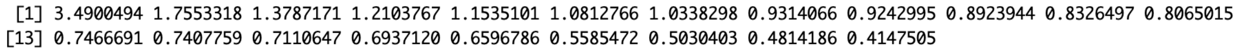




```
       variance.percent cumulative.variance.percent
Dim.1         16.619283                    16.61928
Dim.2          8.358723                    24.97801
Dim.3          6.565319                    31.54333
Dim.4          5.763699                    37.30702
Dim.5          5.492905                    42.79993
Dim.6          5.148936                    47.94887
Dim.7          4.922999                    52.87186
Dim.8          4.435270                    57.30713
Dim.9          4.401426                    61.70856
Dim.10         4.249497                    65.95806
Dim.11         3.964999                    69.92306
Dim.12         3.840483                    73.76354
Dim.13         3.555567                    77.31911
Dim.14         3.527504                    80.84661
Dim.15         3.386023                    84.23263
Dim.16         3.303390                    87.53602
Dim.17         3.141327                    90.67735
Dim.18         2.659748                    93.33710
Dim.19         2.395430                    95.73253
Dim.20         2.292469                    98.02500
Dim.21         1.975002                   100.00000
```

We use the number of dimensions to have at least 70% of variance; hence in this question, we should choose 11 PCA.

At last we can use Kaiser’s rule,



```
 [1] 3.4900494 1.7553318 1.3787171 1.2103767 1.1535101 1.0812766 1.0338298 0.9314066 0.9242995 0.8923944 0.8326497 0.8065015
[13] 0.7466691 0.7407759 0.7110647 0.6937120 0.6596786 0.5585472 0.5030403 0.4814186 0.4147505
```

Kaiser's rule is a guideline used in principal component analysis (PCA) to determine the number of principal components to retain for modeling. It is based on the idea that only the components with eigenvalues greater than or equal to 1 should be retained.

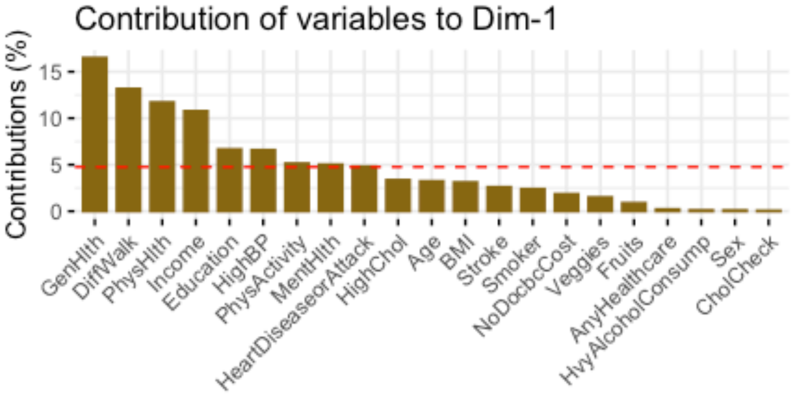
Kaiser's rule states that the number of principal components to retain is equal to the number of components with eigenvalues greater than or equal to 1. This is because eigenvalues represent the amount of variance explained by each component, and components with eigenvalues less than 1 explain less variance than a single original variable.

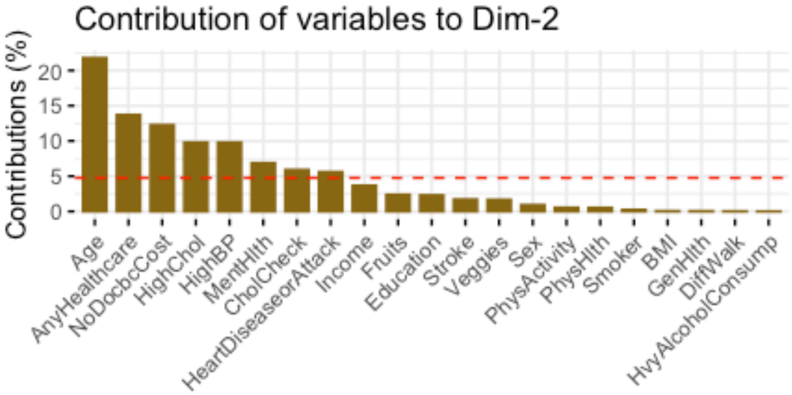
Kaiser's rule is a common method for selecting the number of principal components to retain, but it is not always the best method. In some cases, it may be more appropriate to use other ways, such as the scree plot or cumulative variance plot, to determine the number of components to retain. It is also essential to consider the interpretability of the included features and their relevance to the research question or problem being studied.

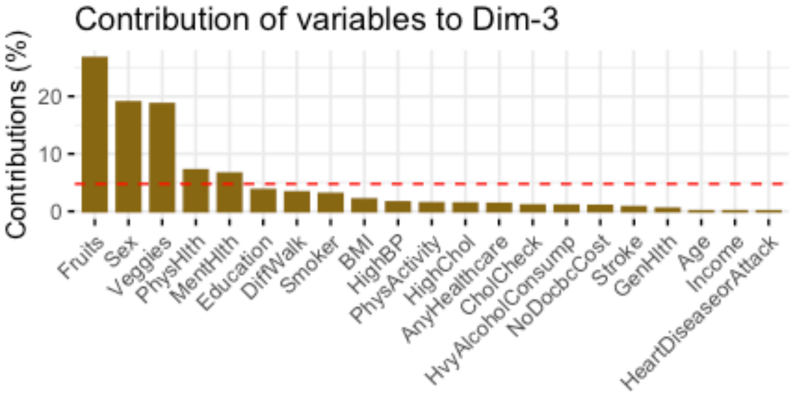
Hence, we use 7 PCA because & all of our dimensions have E.V.>1.

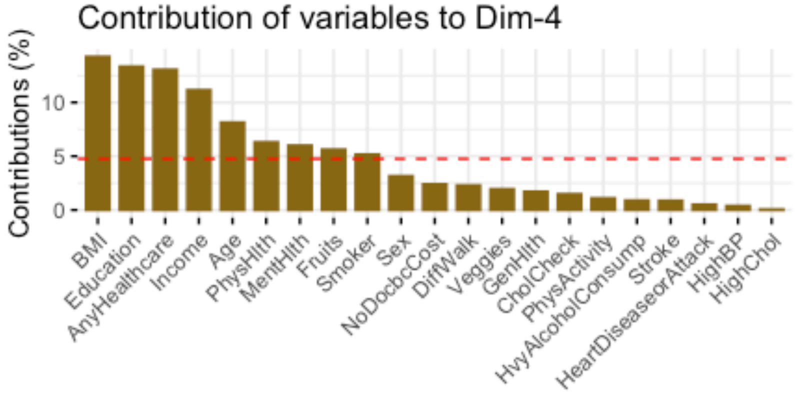
We now know that we should select seven features.

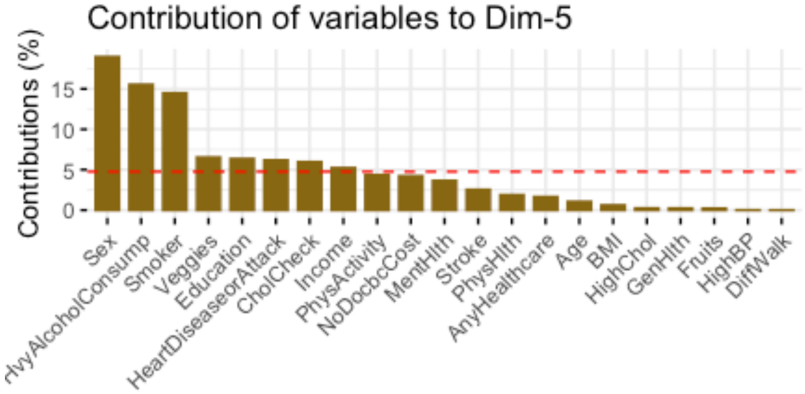
So now we use PCA:

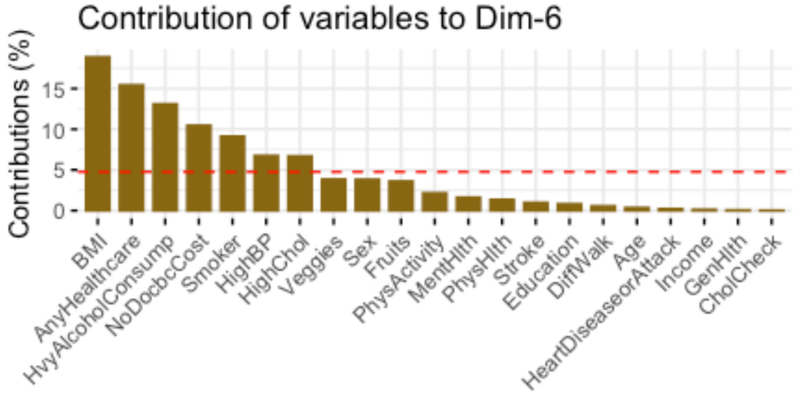

Contribution of variables to Dim-1

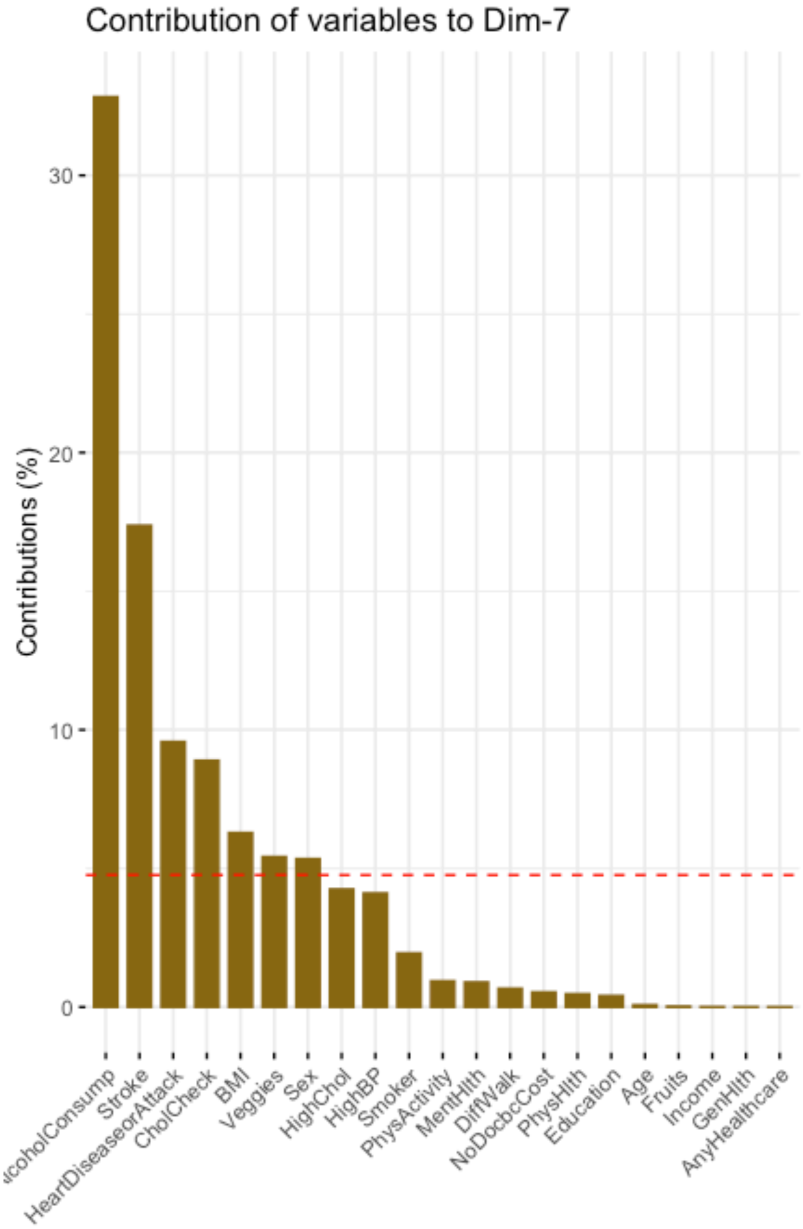

Contribution of variables to Dim-2


Contribution of variables to Dim-3

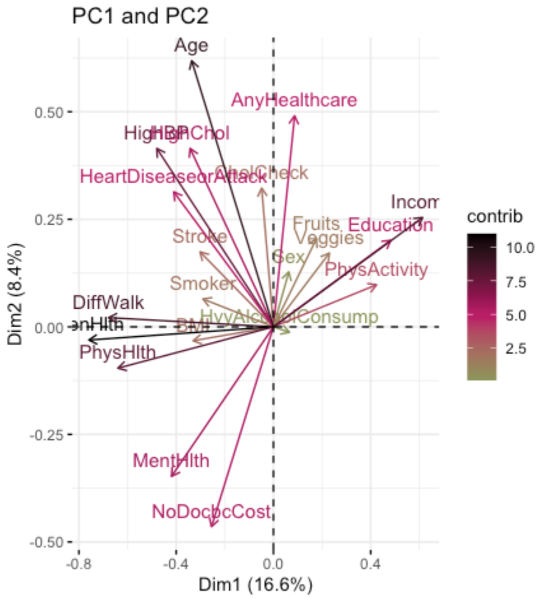

Contribution of variables to Dim-4
Contributions (%)
0
5
10
BMI
Education
AnyHealthcare
Income
Age
PhysHlth
MentHlth
Fruits
Smoker
Sex
NoDocbcCost
DiffWalk
Veggies
GenHlth
CholCheck
PhysActivity
HvyAlcoholConsump
Stroke
HeartDiseaseorAttack
HighBP
HighChol

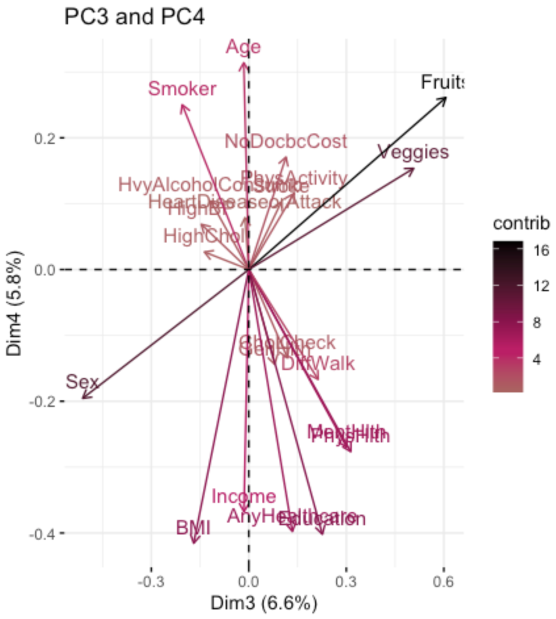

Contribution of variables to Dim-5
Contributions (%)
0
5
10
15
Sex
vyAlcoholConsump
Smoker
Veggies
Education
HeartDiseaseorAttack
CholCheck
Income
PhysActivity
NoDocbcCost
MentHlth
Stroke
PhysHlth
AnyHealthcare
Age
BMI
HighChol
GenHlth
Fruits
HighBP
DiffWalk


Contribution of variables to Dim-6
Contributions (%)
0
5
10
15
BMI
AnyHealthcare
HvyAlcoholConsump
NoDocbcCost
Smoker
HighBP
HighChol
Veggies
Sex
Fruits
PhysActivity
MentHlth
PhysHlth
Stroke
Education
DiffWalk
Age
HeartDiseaseorAttack
Income
GenHlth
CholCheck



And for the first and second components we also have the:



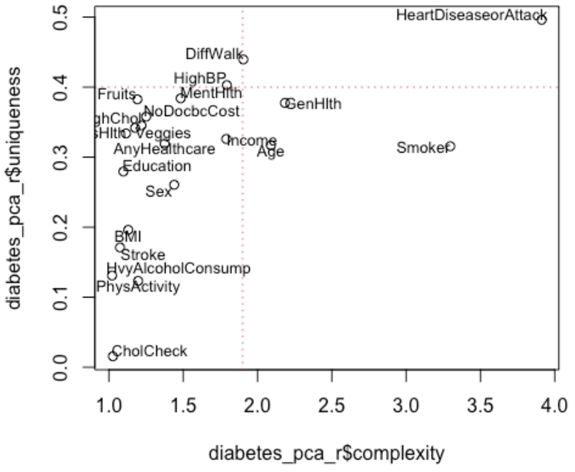


So we can say that if we want to choose one variable, it's better to choose Genhealth. (or age).

Second based on contribution and also considering variance between some variables, we choose our seven variables like this:

"Age/PhysActivity/Sex/HvyAlcoholConsump/BMI/HighBP/Stroke"

We also considered Complexity and Uniqueness and canceled some variables here too:



We use simple logistic regression for this problem.

4. As we said, we chose our seven features for more straightforward and better classification.

What we used was a multinom model for classification.

It would be best to consider that using an unbiased dataset for training is better because our model won't be biased toward a specific group.

You can't use a validation dataset because you don't need it.

The best way is to split data 80/10/10 % or 80/20%.

The best way to show data is also a ROC chart because it's more important not to have FalseNegative in our model than high accuracy.

5.As we trained a good model with acceptable result we can say by asking only 7 questions we can have found about this probability of patient diabetes.