Analysis on Breast Cancer and Possible Factors

Multivariate Analysis Term Project

Ekinsu ÇİÇEK Mehmet Ali ERKAN

OUTLINE

- Introduction to the dataset
- Analysis on Research Question 1
- Analysis on Research Question 2
- Analysis on Research Question 3
- Conclusion

Breast Cancer Data Set

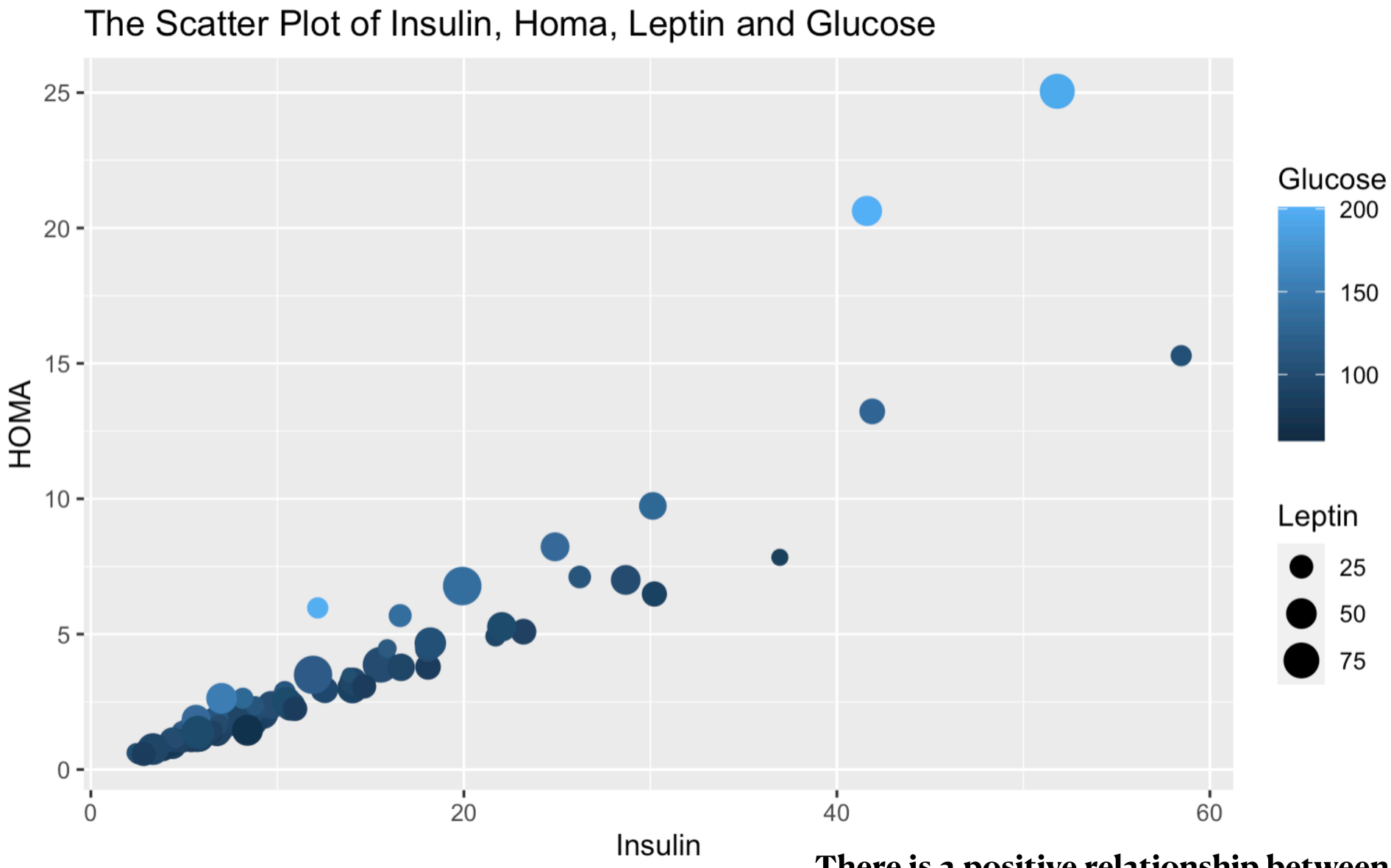
	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
1	48	23.50	70	2.70	0.46	8.80	9.70	7.99	417.11	1
2	83	20.69	92	3.11	0.70	8.84	5.42	4.06	468.78	1
3	82	23.12	91	4.49	1.00	17.93	22.43	9.27	554.69	1
4	68	21.36	77	3.22	0.61	9.88	7.16	12.76	928.22	1
5	86	21.11	92	3.5	0.80	6.69	4.81	10.57	773.92	1
6	49	22.85	92	3.2	0.73	6.83	13.67	10.31	530.41	1

- 116 observation with 10 variables
- 36 outliers in the data.

Variable	Information	Type
Age	years	integer
BMI	kg/m2	numeric
Glucose	mg/dL	integer
Insulin	µU/ml	numeric
HOMA		numeric
Leptin	ng/mL	numeric
Adiponectin	µg/mL	numeric
Resistin	ng/mL	numeric
MCP.1	pg/dL	numeric
Classification	1 = healthy 2 = patients	Factor levels('1','2')

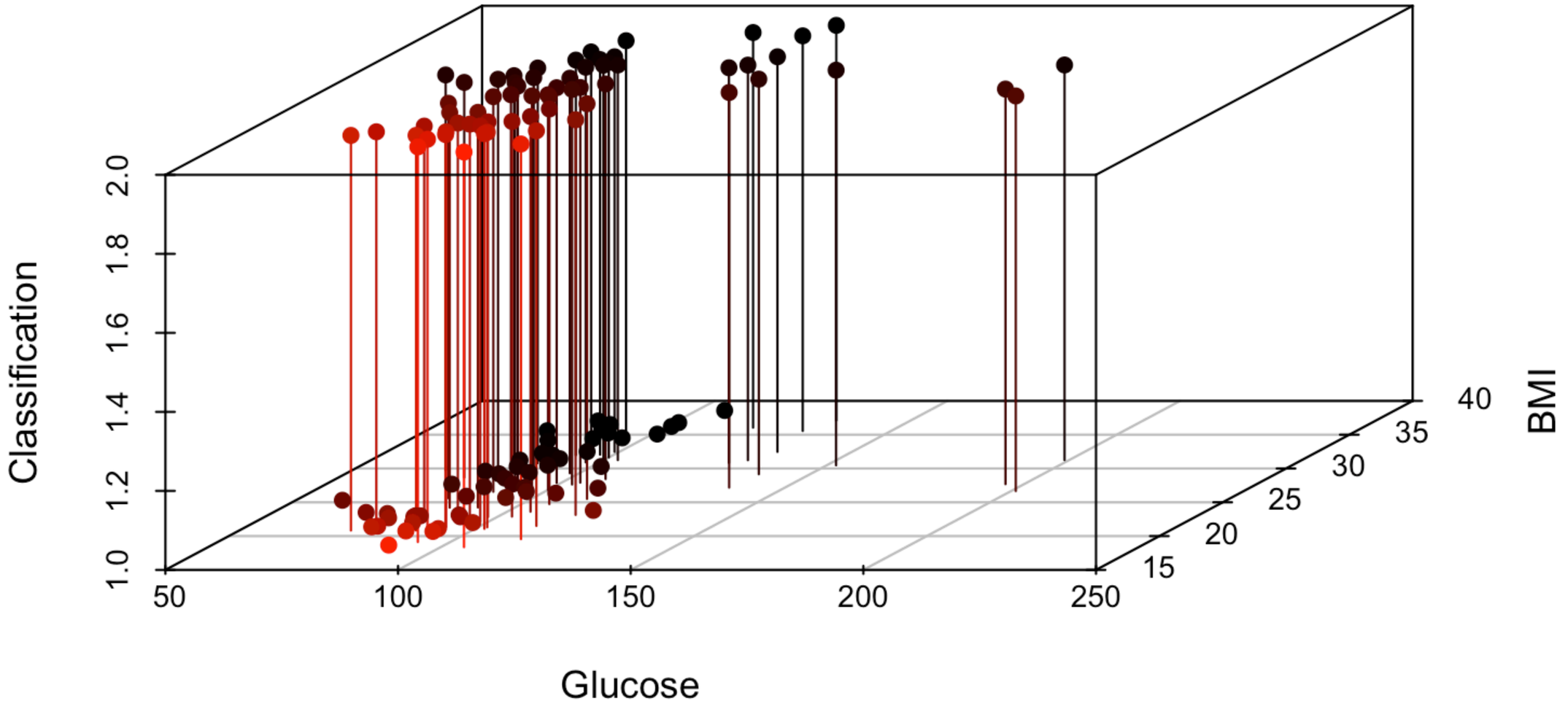
	Age	*BMI*	*Glucose*	*Insulin*	*HOMA*	*Leptin*	*Adiponectin*	*Resistin*	*MCP.1*	*Classification*
Min	24.0	18.37	60.00	2.43	0.46	4.31	1.65	3.21	45.84	1.00
1st Qu.	45.0	22.97	85.75	4.35	0.91	12.31	5.47	6.88	269.98	1.00
Median	56.0	27.66	92.00	5.92	1.38	20.27	8.35	10.82	471.32	2.00
Mean	57.3	27.58	97.79	10.01	2.69	26.61	10.18	14.72	534.65	1.55
3rd Qu.	71.0	31.24	102.00	11.18	2.85	37.37	11.81	17.75	700.09	2.00
Max.	89.0	38.58	201.00	58.46	25.05	90.28	38.04	82.10	1698.44	2.00

The Scatter Plot of Insulin,Homa, Leptin and Glucose

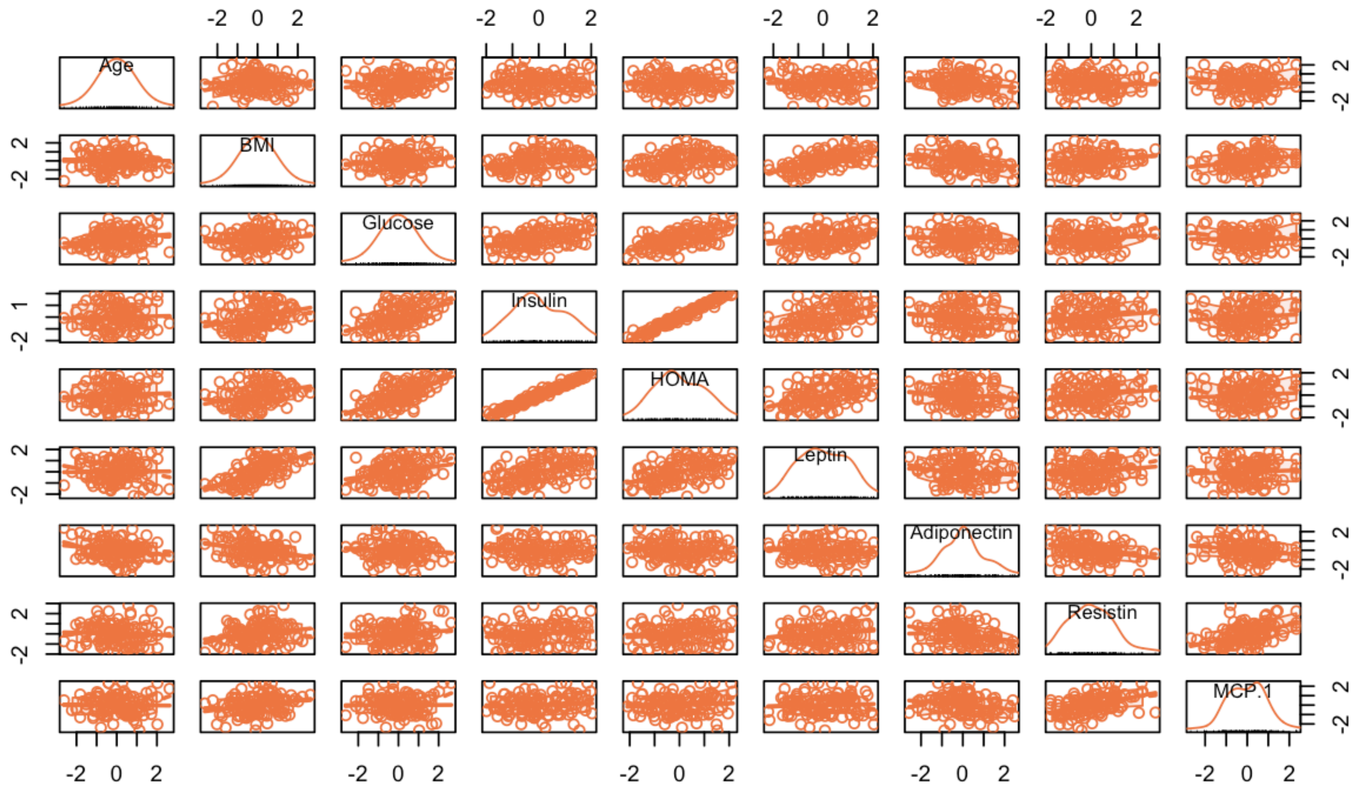


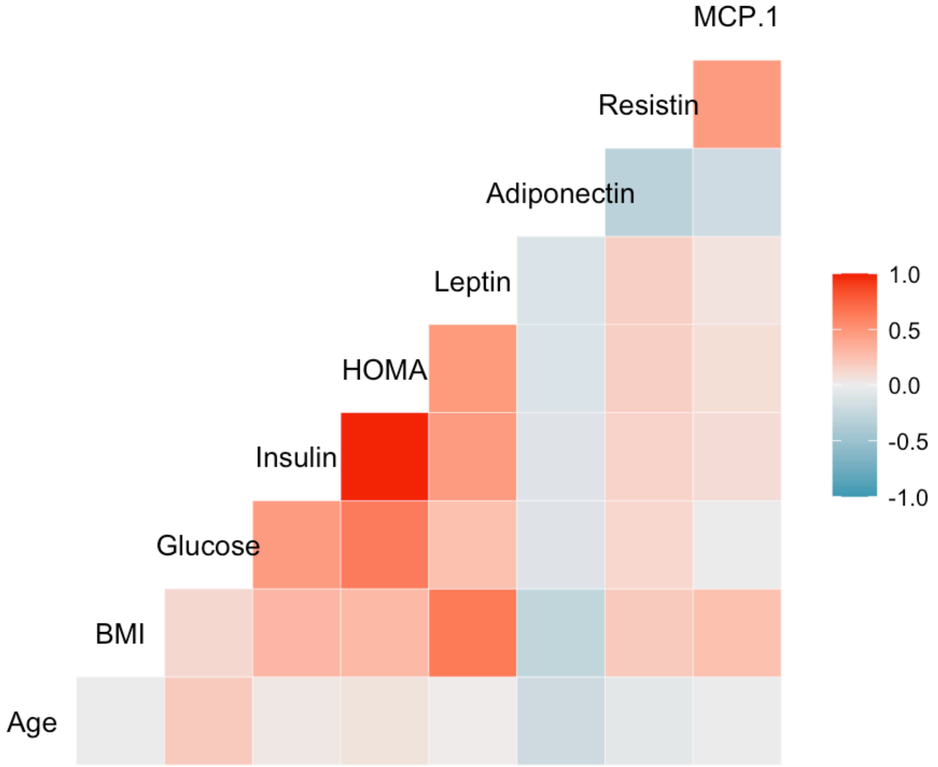
There is a positive relationship between variables

Relationship between Glucose - BMI - Classification



Is there a relationship between Insulin and HOMA level on blood and AGE, BMI and other hormones?





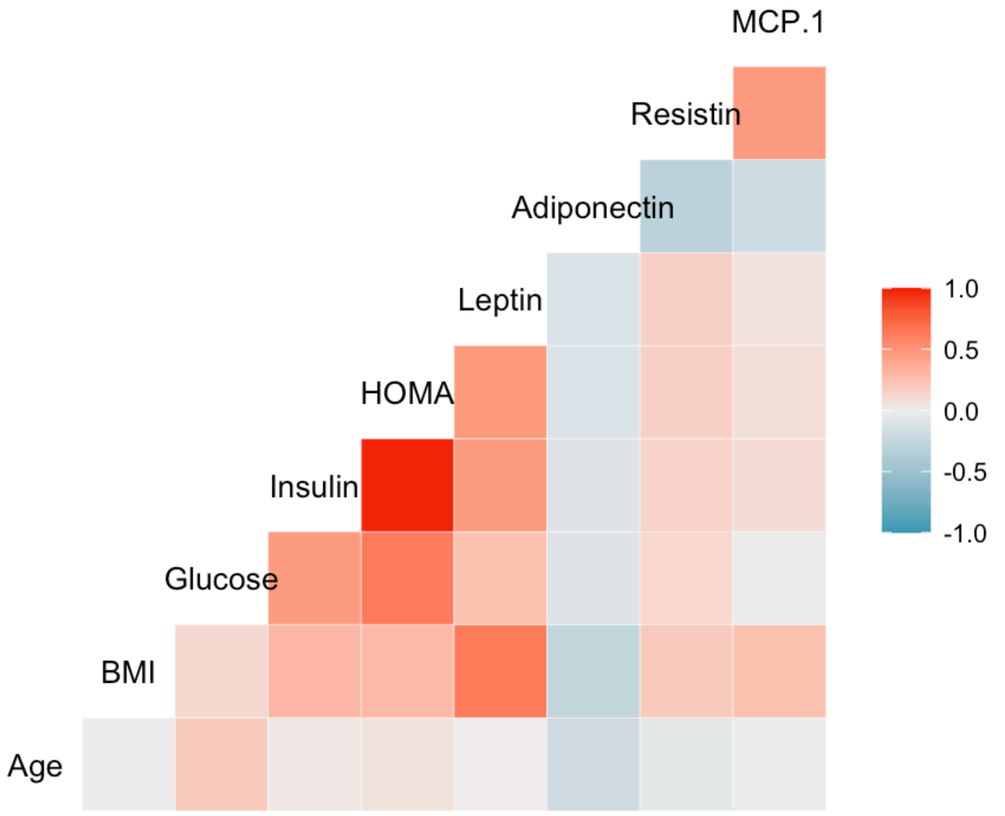
Our dimensions are 116 and 9, we applied principle component analysis before creating a model.

Scaling

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
-0.400	-0.532	-2.120	-1.714	-2.094	-1.219	0.2235	-0.514	-0.141
1.594	-1.672	-0.076	-.1378	-1.211	-1.214	-0.728	-1.490	0.034

Since the scale of the variables are different, data was scaled before PCA.

Correlations of the variables seem good for principle component analysis.

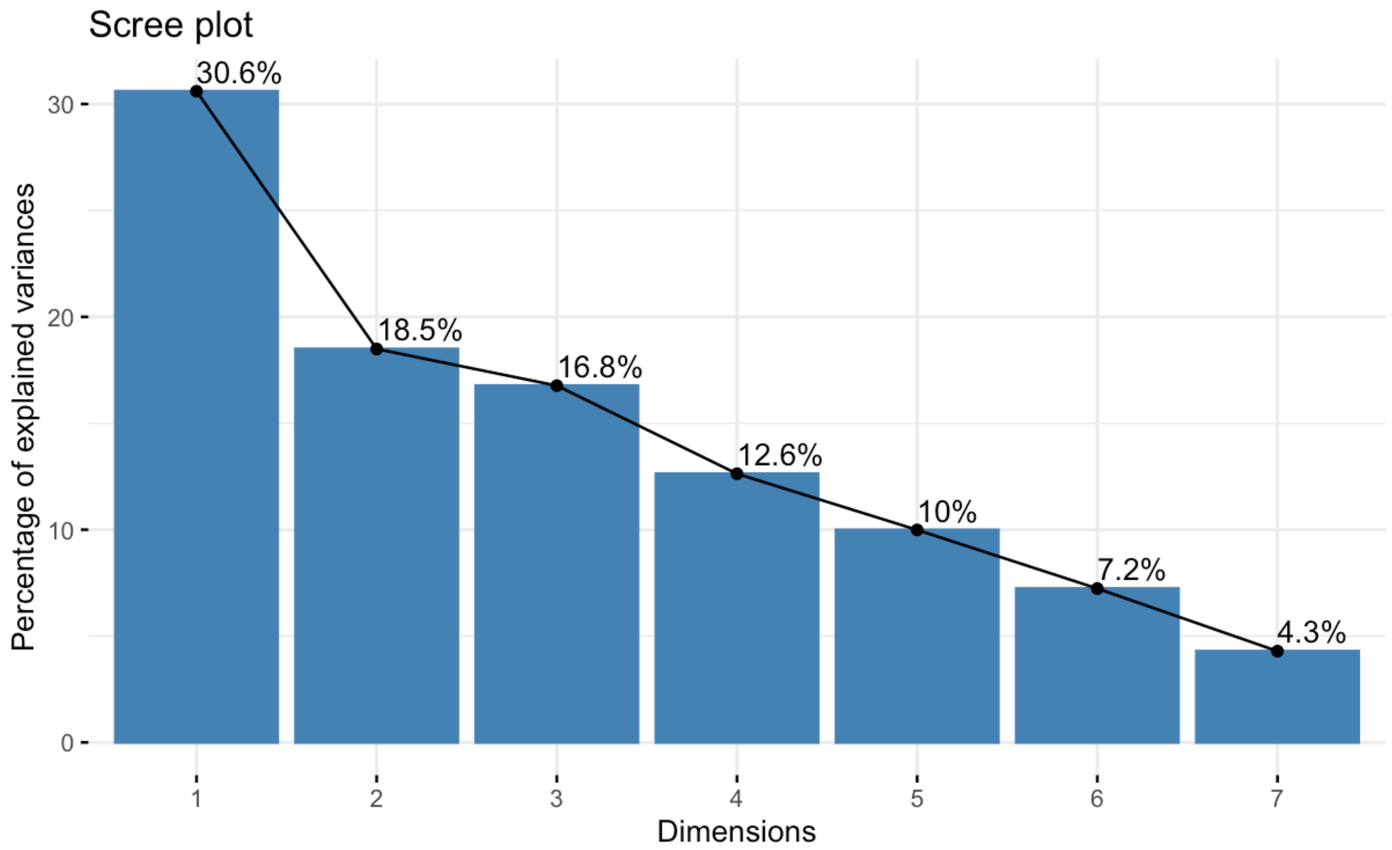


Principal Components and Eigenvalues

Eigenvalues
1.4635859
1.1376775
1.0833787
0.9401669
0.8361664
0.7116267
0.5480844

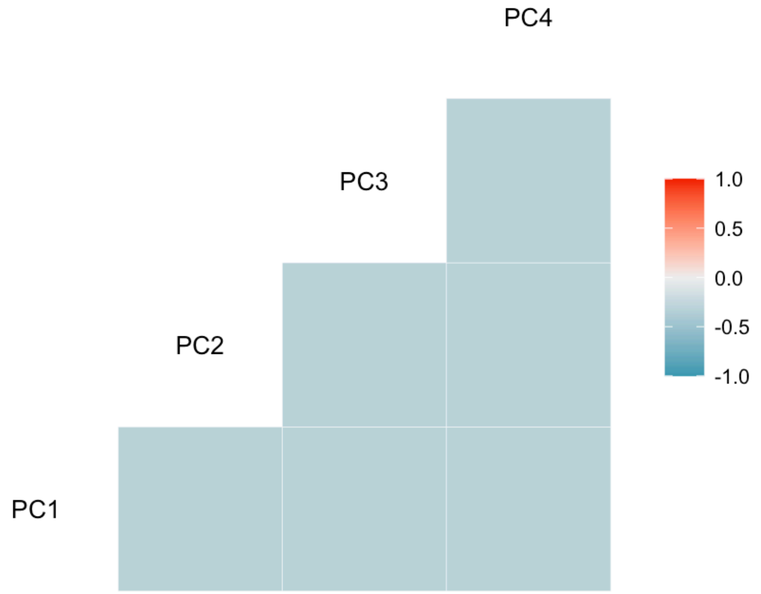
The first four components nearly explains 80% variability of the data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard Deviation	1.464	1.1377	1.0834	0.9402	0.83617	0.71163	0.54808
Proportion Of Variance	0.306	0.1849	0.1677	0.1263	0.09988	0.07234	0.04291
Cumulative Proportion	0.306	0.4909	0.6586	0.7849	0.88474	0.95709	1.00000



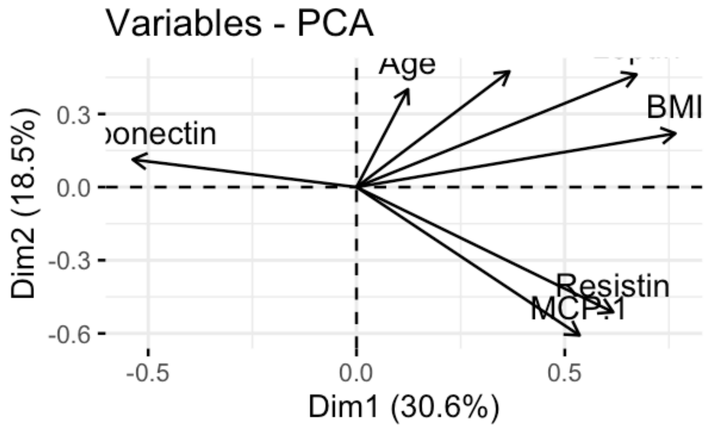
Scree plot shows that first four component are suitable for explaining variability in the data.

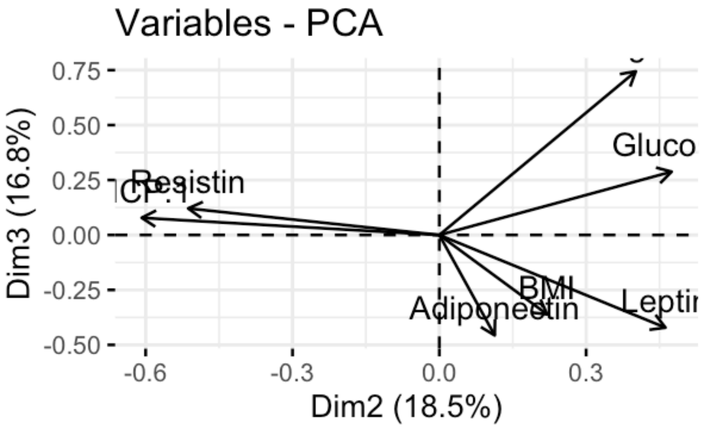
Correlation Plot of Principal Components

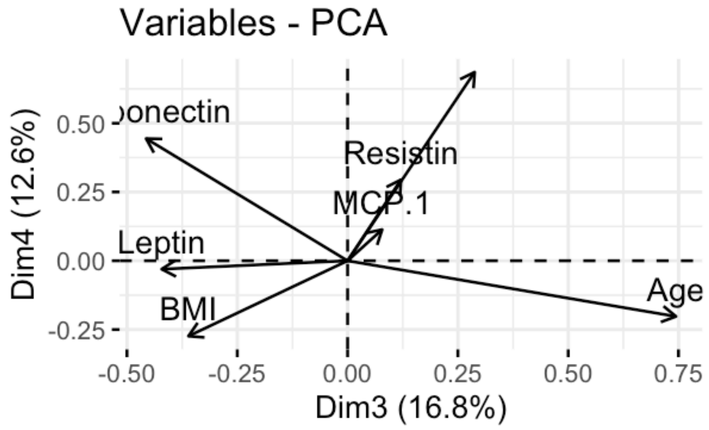


Correlation of four components shows that they are linear independent.

	PC1	PC2	PC3	PC4
Age	0.1234575	0.4013896	0.74425187	-0.20239975
BMI	0.7653600	0.2202011	-0.36066473	-0.27393188
Glucose	0.3668419	0.4746081	0.28793469	0.68563259
Leptin	0.6719581	0.4620263	-0.42135976	-0.03041939
Adiponectin	-0.5371351	0.1131741	-0.45666194	0.44421208
Resistin	0.6161955	-0.5131530	0.12084092	0.29446678
MCP.1	0.5354946	-0.6081549	0.07827094	0.11338845







The Relationship Between Components and Variables

Principle Component Regression

Insulin = -5.157e-17 + 2.894e-01 PC1 + 2.546e-01 PC2 - 6.256e-02 PC3 + 2.698e-01 PC4	
Residual Standard Error : 0.8318	MSE : 0.6735252
Multiple R- Squared : 0.3322	Adjusted R- Squared : 0.3082
F – Statistics : 13.81	P – Value: 5.588e-09

HOMA = 3.032e-16 + 3.128e-01 PC1 + 3.113e-01 PC2 -5.496e-03 PC3 + 3.962e-01 PC4	
Residual Standard Error : 0.8318	MSE : 0.5332795
Multiple R- Squared : 0.3322	Adjusted R- Squared : 0.3082
F – Statistics : 13.81	P – Value: 5.588e-09

Multivariate Linear Regression

Insulin = -0.0001605 - 0.0386830 Age + 0.0125844 BMI + 0.3735138 Glucose + 0.3591940 Leptin - 0.0088957 Adiponectin + 0.0071172 Resistin + 0.0588322 MCP.1	
Residual Standard Error : 0.8329	MSE : 0.6597791
Multiple R- Squared : 0.3484	Adjusted R- Squared : 0.3062
F – Statistics : 8.251	P – Value: 5.894e-08

HOMA = -0.0002302 - 0.0450229 Age + 0.0196093 BMI + 0.5482072 Glucose + 0.3134629 Leptin - 0.0094384 Adiponectin + 0.0047811 Resistin + 0.0542865 MCP.1	
Residual Standard Error : 0.7323	MSE : 0.5131732
Multiple R- Squared : 0.4963	Adjusted R- Squared : 0.4637
F – Statistics : 15.2	P – Value: 1.046e-13

All models are significant but R-Squared values of them very low. Also, mean standard error of the models are high.

Glucose - BMI one sample T2-test

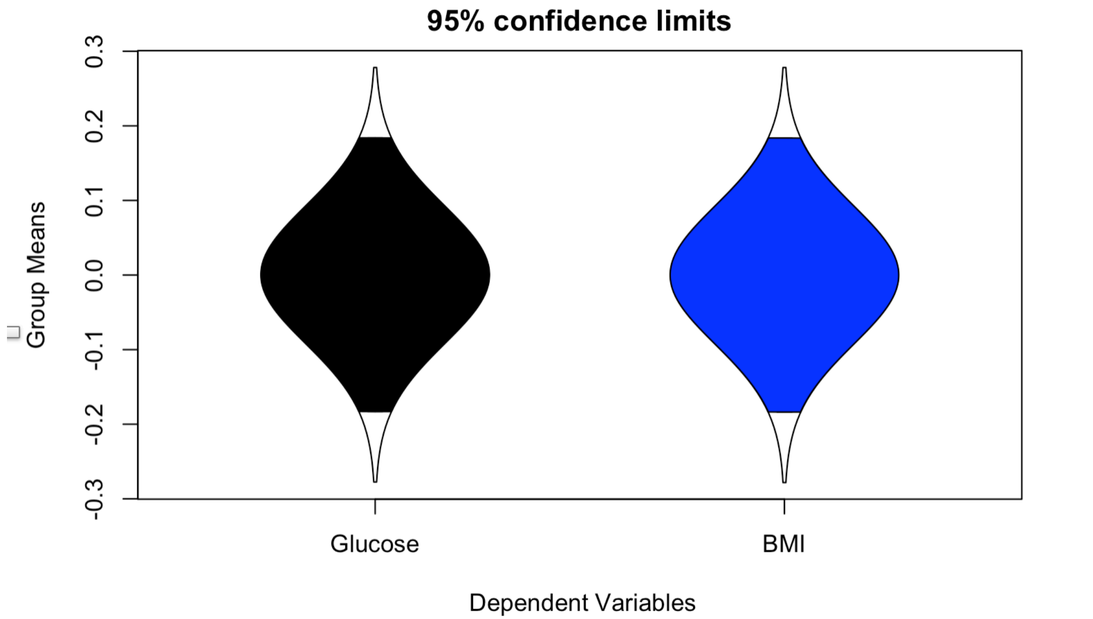
Normality Check

Test	statistics	statistics	result
Mardia Skewness	0.59	0.96	yes
Mardia Skewness	-0.96	0.33	yes

Hotelling's one sample T2-test

data	Glucose-BMI
Value	T.2=4915.5, df1=2, df=114, p < 2e-16

Response Matrix



Does the classification have a significant effect on Glucose and BMI variables?

Factor 1	Factor 2
52	64

Classification	variable	statistics	p
1	BMI	0.99	0.96
1	Glucose	0.99	0.98
2	BMI	0.99	0.96
2	Glucose	0.99	0.94

data follows univariate normality

Box's M-test for Homogeneity of Covariance Matrices

data	Glucose-BMI
Value	Chi-Sq= 2,18, df=3, p = 0.53

variance-covariance matrices are equal

Hotelling's two sample T2-test

data	Glucose-BMI by Classification
Value	T.2= 17,308, df=2, p = 2.77e-07

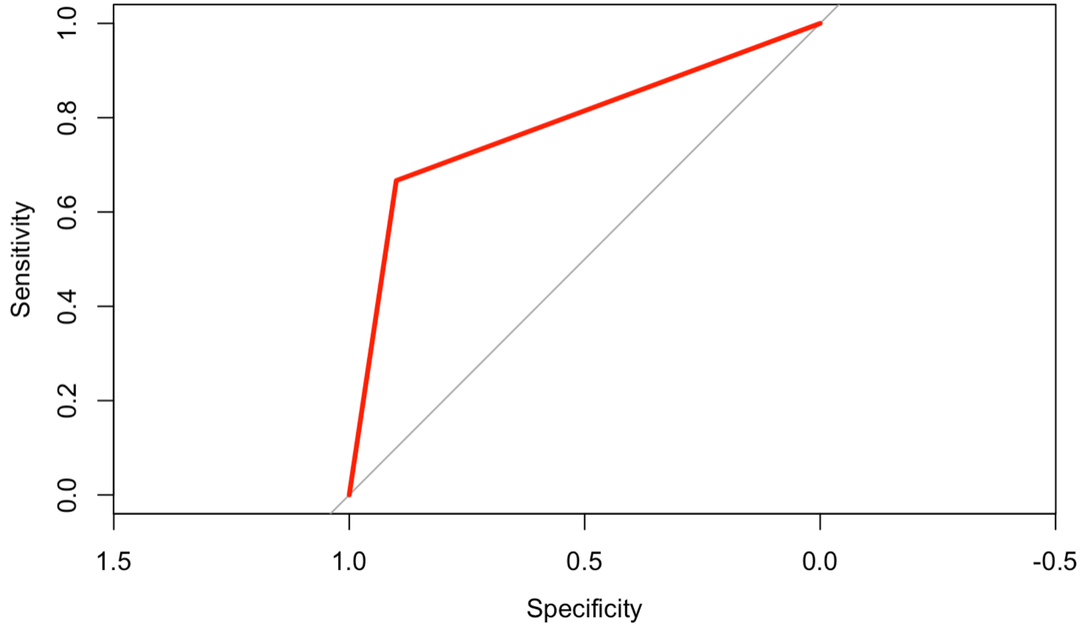
How accurate is the classification for detecting the disease?

Accuracy	**0.77**
95%CI	(0.54, 0,92)
No Information Rate	(0.54)
P value [Acc>Nır]	(0.02)
Kappa	0.55
Mcnemar's Test P-Value	0.37
Sensitivity	0.90
Specificity	0.66
Pos Pred Value	0.69
Neg Pred Value	0.88
Prevalence	0.45
Detection Rate	0.40
Detection Prevalance	0.59
Balanced Accuracy	0.78

Confusion Matrix

Prediction	Reference 1	Reference 2
1	9	4
2	1	8

ROC CURVE




Conclusion