The study of factors in heart diseases, the case of Albania

Frida Zisko, Agresa Qosja

Abstract— This article presents the study of factors in heart disease. The data were obtained in a private institution in Albania and the causes from which patients suffer were studied. Statistical study was done using the SPSS program. The obtained results are of interest to medical researchers and data analysts.

Index Terms – Statistics, Cardiovascular disease, SPSS, Linear regression, Chernoff faces, Cardiologic factors.

1 INTRODUCTION

One of the problems of society nowadays, which has been worrying for doctors and individuals, are heart diseases and specifically coronary diseases that affect every age and gender. In this study, we have focused more on the factors possessed by patients presented for visits, analysis, coronarography, in a private health institution. The time period in which the data were collected is six months.

Coronarography is a procedure in which the coronary vessels are observed in detail, precisely those vessels that feed the heart with blood, the blockage of which can bring consequences up to fatal for the patients. A heart attack is one of the complications of the blockage of these vessels. A heart attack is a consequence of the blockage of the blood vessels that supply the heart muscle with blood. As a result, serious disorders appear in the work of the heart, the causes of which are various. These causes include: heredity, obesity, hypertension, high blood cholesterol, diet and lifestyle, smoking, etc.

The purpose of the study is to look at the direct or indirect impact of these factors on heart disease.

The data were collected in a private hospital in Albania. The study is done on 200 patients, of which 151 are men and 49 are women. The variables on which the study was carried out are age, gender, smoking, AFP, HTA, FBG, TG, HbA1c, COL, LDL_C, HDL_C, CAD, ANGINA.

The variable	The description
AGE	The age of patient
SEX	The gender of patient
AFP	The presence of alpha feto protein in the patient
FUMATOR	If you smoke or not
HTA	If there are problems with hypertension
FBG	Amount of fasting blood glucose
TG	The amount of triglycerides
HbAlc	Hemoglobin, type A1c
COL	The amount of cholesterol
LDL_C	The amount of low density lipo-protein
HDL_C	The amount of high density lipo-protein
CAD	Coronary artery disease
ANGINA	Angina pectoris

Table 1.1 Description of variables

Application of linear regression

1.1 Indicators of coronary artery disease.

Coronary artery disease (CAD)

Coronary artery disease is the basis of heart attacks. The heart is the muscle that is surrounded and fed by blood vessels. These blood vessels for various reasons begin and narrow from plaques that originate from fats (therefore it is given importance in the study). A heart attack develops gradually over time. As we can see and from the data collected for our study, we have 3 types of blood vessel blockages:

1. Single - when only one blood vessel is blocked, while the others are open.
2. Multiple - when more than one blood vessel is blocked.
3. No - when no blood vessel is blocked.

If only one is blocked, it is enough to narrow or create infarction in a certain part of the heart.

Frida Zisko, Mediterranean University of Albania
fridazisko@umsh.edu.al
Agresa Qosja, Metropolitan Tirana University
aqosja@umt.edu.al

Figure 1.1. Heart

Smoking

The use of tobacco is a risk factor for the development of coronary heart disease. Despite the well-known damage of smoking on heart health, the knowledge of certain groups of the population about smoking, as one of the main causes of cardiovascular diseases, is scarce.

		Coronary Artery Disease			Total
		Multiple	No	Single	
Smoker	No	49	29	43	121
	Yes	25	14	40	79
Total		74	43	83	200

Table 1.2 Contingency table between Smoking and CAD

From the contingency table obtained from SPSS, we see that of the 200 patients taken in the study, 39.5% (79) smoke while 60.5% do not (121). Of those who consume, 65 individuals, i.e., about 82.3%, narrowing of blood vessels (40 have only one blocked blood vessel while 25 have several blocked blood vessels), while of those who do not consume, 92 individuals, i.e., about 76%, have narrowing of blood vessels. theirs.

So, we see that individuals who smoke are more predisposed to have narrowing of the blood vessels and therefore the presence of coronary artery disease. We thus confirmed that smoking is a factor in heart attacks.

1.1.2 Gender of patients

Out of 200 people, 151 of them are men and 49 are women. We see that the number of people who suffer more from heart diseases are men, of which 24 of them do not have blocked arteries, while 58 have several and 69 only one blocked artery. Of the 49 women, 19 have no blocked arteries while 16 have several and 14 only one blocked artery.

Of the total number of patients, 63.5% of men and 16% of women have at least one blocked artery. This result showed us that we have male dominance in coronary disease.

		Coronary Artery Desease			Total
		Multiple	No	Single	
Sex of the Pacient	Male	58	24	69	151
	Female	16	19	14	49
Total		74	43	83	200

Table 1.3 Contingency table for Gender and CAD

Age of patients

Statistics		
Age of the Pacient		
N	Valid	200
	Missing	0
Median		61.00
Mode		60
Minimum		30
Maximum		82

Table 1.4 Statistics for age

As we can see from table 1.4, the age ranges from 30 to 82 years. The average age is 61 years and the most repeated value is 60 years. So middle age should be more careful.

1.2 Multiple Linear Regression

Models with more than one explanatory variable are called multiple linear regression models.
In the multiple linear regression model, the data must: have a normal distribution, have a linear relationship between them, the mean of the error is zero, constant variance, no autocorrelation, no multiple relationships between independent variables.
To show which of the factors: age, gender, presence of alpha feto protein, smoking, hypertension, glycemia (sober), triglycerides, hemoglobin, cholesterol, low density lipoprotein and high-density lipoprotein affects coronary artery disease, we build the model of linear regression.

Statistics

		Age of the Pacient	Sex of the Pacient	Alfa Feto-Protein	Smoker	Hipertension	Fasting Blood Glucose	Tryglicerydes	Hemoglobin type A1c	Colesterol	Low Density Lipoprotein	High Density Lipoprotein	Coronary Artery Desease
N	Valid	200	200	200	200	200	200	200	200	200	200	200	200
	Missing	0	0	0	0	0	0	0	0	0	0	0	0

Table 1.5 Number of inclusion of variables in the model

The table above is the first table we get in SPSS. This table gives us information about the number of cases that are included in the analysis of the model and those that are not considered in the model.
In our specific case we have no missing data, in this way our model is built based on all the data we have taken into consideration.

The multiple regression model has the form:
$y = \beta_0 + \beta_1 x_1 + ... + \beta_i x_i + \varepsilon$

Hypothesis H_0 in the multiple regression model indicates that all regression coefficients are equal to zero.
($H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$).

The HA hypothesis states that at least one βi is different from zero.
To statistically test the significance of the parameters separately, the t-test is used, and to test whether the model is significant as a whole, the F-test is used.

Application of the method of elimination of variables (Backward Elimination)
This method will start from the full model until it reaches the simple model. Regress the CAD variable on all the regressors under consideration. Invalid will be those regressors that have the value of Sig. > 0.1 but the one with the smallest t-statistic value will be excluded. The regression will be calculated again until the most important independent variables remain in the model.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.393[a]	.155	.105	.710	.155	3.125	11	188	.001	
2	.393[b]	.154	.110	.709	.000	.049	1	188	.826	
3	.392[c]	.154	.114	.707	-.001	.127	1	189	.722	
4	.391[d]	.153	.118	.705	-.001	.152	1	190	.697	
5	.390[e]	.152	.121	.704	-.001	.258	1	191	.612	
6	.388[f]	.150	.124	.703	-.002	.399	1	192	.528	
7	.384[g]	.148	.126	.702	-.003	.583	1	193	.446	
8	.379[h]	.144	.126	.702	-.004	.837	1	194	.361	
9	.373[i]	.139	.126	.702	-.005	1.067	1	195	.303	2.161

Table 1.6 Models of the Backward Elimination method

a. Predictors: (Constant), High Density Lipoprotein, Smoker, Cholesterol, Alpha Feto-Protein, Hypertension, Fasting Blood Glucose, Age of the Patient, Sex of the Patient, Triglycerides, Hemoglobin type A1c, Low Density Lipoprotein
b. Predictors: (Constant), High Density Lipoprotein, Smoker, Cholesterol, Alpha Feto-Protein, Hypertension, Fasting Blood Glucose, Age of the Patient, Sex of the Patient, Hemoglobin type A1c, Low Density Lipoprotein

c. Predictors: (Constant), High Density Lipoprotein, Smoker, Cholesterol, Alpha Feto-Protein, Hypertension, Age of the Patient, Sex of the Patient, Hemoglobin type A1c, Low Density Lipoprotein
d. Predictors: (Constant), High Density Lipoprotein, Smoker, Cholesterol, Alpha Feto-Protein, Age of the Patient, Sex of the Patient, Hemoglobin type A1c, Low Density Lipoprotein

e. Predictors: (Constant), High Density Lipoprotein, Smoker,

Alfa Feto-Protein, Age of the Patient, Sex of the Patient, Hemoglobin type A1c, Low Density Lipoprotein

e. Predictors: (Constant), High Density Lipoprotein, Alpha Feto-Protein, Age of the Patient, Sex of the Patient, Hemoglobin type A1c, Low Density Lipoprotein

g. Predictors: (Constant), High Density Lipoprotein, Age of the Patient, Sex of the Patient, Hemoglobin type A1c, Low Density Lipoprotein

h. Predictors: (Constant), Age of the Patient, Sex of the Patient, Hemoglobin type A1c, Low Density Lipoprotein

i. Predictors: (Constant), Age of the Patient, Sex of the Patient, Hemoglobin type A1c

j. Dependent Variable: Coronary Artery Disease

The Durbin Watson coefficient is used to test for autocorrelation. Our value is close to 2 and indicates no autocorrelation. Durbin Watson values are preferably 1.5 to 2.5. Positive autocorrelation means that the standard error of the coefficient b is very small.

From the table in the section standardized beta coeffiecients shows the order of importance of independent variables without considering the sign of Beta.
The variable with the highest Beta value, i.e. HbA1c, is the most important independent variable, followed by age and finally the patient's gender.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
9		B	Std. Error	Beta			Lower bound	Upper bound
	(Constant)	-.514	.356		-1.446	.150	-1.216	.187
	Age of the Patient	.017	.005	.230	3.407	.001	.007	.027
	Sex of the Patient	-.377	.118	-.216	-3.199	.002	-.609	-.144
	Hemoglobin type A1c	.106	.031	.231	3.459	.001	.046	.167

Table 1.7 Table of coefficients for the final model

The final model will have the form:
CAD = -0.514 + 0.106 *HbA1c + 0.017*Age - 0.377* Sex
The meaning of this is that: The smaller the value of the sugared hemoglobin and the age, the lower the value of CAD will be for each of the sexes. It will be even smaller for women. If they will be zero, then the CAD value will decrease by 0.514 units.

As we see from Table 1.6, 9 models are built. In the final model, 13.9% of the variation in the dependent variable is explained by the independent variables. The coefficient of determination (R square) shows that 13.9% of CAD is explained by HbA1c, Age and Gender of the patient. Adjusted R square shows that 12.6% of CAD is explained by those three but only for the variables that are related to the model.

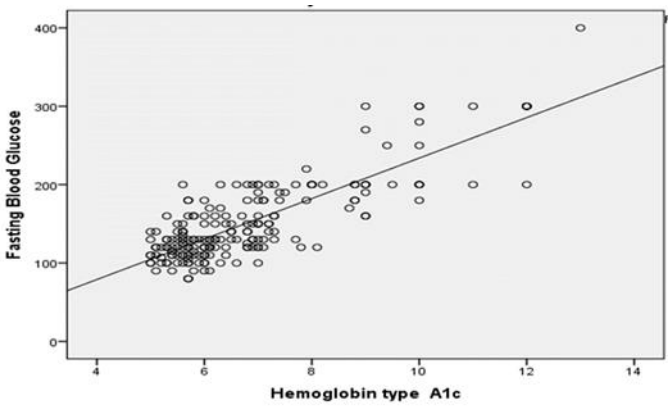

Figure 1.2 Regression Line for HbA1c and FBG

Figure 1.2 shows that there is a linear relationship between glycated Hemoglobin and Glycemia. The amount of glycemia will affect the level of Hemoglobin, which on the other hand we said is the most important influencing factor in coronary artery disease. It is suggested that its amount be kept as low as possible. Linear R^2 is 0.672.

Model		Sum of squares	df	Mean square	F	Sig.
9	Regression	15.627	3	5.209	10.572	.000[j]
	Residual	96.568	196	.493		
	Total	112.195	199			

Table 1.8 ANOVA table

The ANOVA table is useful for testing the significance of the model as a whole. The F value in the table of 10.572 indicates that our model is significant at every level (Sig. = .000).

	Min	Max	Mean	Std. Deviation	N
Predicted Value	0.45	1.89	1.16	.280	200
Residual	-1.470	1.400	.000	.697	200
Std.Predicted Value	-2.500	2.631	.000	1.000	200
Std. Residual	-2.094	1.995	.000	.992	200

Table 1.9 Statistics on Waste

Table 1.9 presents statistics for the remaining values. We have the predicted values, the remainder and both of these standardized. Standardization presents times the predicted value with the standard deviation. These values are known as Pearson residuals, their mean is 0 and standard deviation is 1. We see that the minimum value predicted for coronary artery disease (coronary artery disease) is 0.45 while the maximum is 1.89. These standardized values are at least -2,500 and a maximum of 2,631.

CHAPTER 2

Projection of data in subspaces

2.1 Graphic presentation of data

For the graphical presentation of our data, we used Chernoff faces.

This type of representation is based on 1 correspondence of each variable with 1 characteristic from the face of 1 person. It serves us to classify individuals into groups. Since we said that the age most affected by coronary diseases is around 60, we have built faces for patients between the ages of 59 and 61.

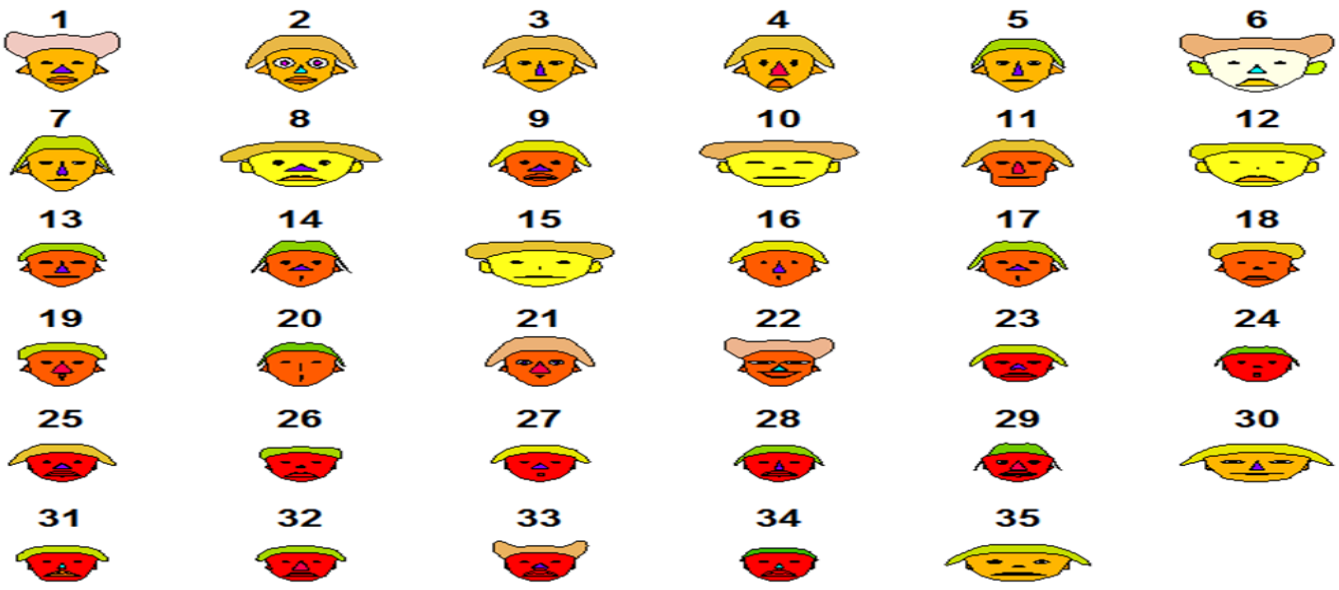

Figure 2.1 Chernoff faces for 35 patients

Variable values:
face length-AGE, face width-SEX, face structure-AFP, mouth length-smoker, mouth width-HTA, smile-FBG, eye length-TG, eye width-HbA1c, hair length- COL, hair width-LDL-C, hair style-HDL-C, nose length-CAD, nose width-ANGINA, ear width-AGE, ear length-SEX.

We see that the most similar individuals are 14 and 17, whose characteristics are:
14: 60,0,0,0,0,115,180,5.8,200,130,30,1,2
17: 60,0,0,0,0,120,140,6.1,200,150,35,1,2
2.2 Principal components analysis

The method of principal components is a method that divides the correlation or covariance matrix into sets of components with the same number as the number of variables involved. An advantage of this method is the simplicity with which we can increase the number of dimensions by adding components of young people.

The purpose of PCA is the projection of a multidimensional set, consisting of quantitative data, into a subset of smaller dimensions. The method is used for cases where we want to identify extreme data or sets of data that differ between them.

Application of the PCA method
For our problem, an empirical analysis was made for CAD of a patient using the variables that most affect CAD. The data set consists of 200 patients, where the dependent variable is CAD and the predictor variables. 7 quantitative variables were obtained from the database depending on which the study was done.
To begin with, for the group of independent variables, we have obtained some numerical indicators such as dispersion, average, etc., where we see a variety of values for different variables.

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Fasting blood glucose	200	80	400	151.01	51.306	2632.261
Tryglicerydes	200	10	600	162.39	95.349	9091.524
Age of the pacient	200	30	82	61.32	10.210	104.249
Hemoglobin type-A1c	200	5.0	13.0	6.790	1.6283	2.652
Low density lipo-protein	200	40	300	138.16	45.578	2077.371
Hight density lipo-protein	200	20	550	46.94	37.551	1410.072
Colesterol	200	10	400	215.50	57.777	3338.191
Valid N (listwise)	200					

Table 2.1 Numerical characteristics for independent variables

What we seem to have in mind when we use the PCA method is the fact that the data must be standardized or we will work with correlation tables as in the case below. In table 2.2 we see the correlation values between the independent variables.

There are marked with ** the most important correlations (Pearson Correlation) according to the confidence level (Sig. (2-tailed) <0.01). The table shows the values of the correlation coefficients between the 7 independent variables.

Correlations

		Age of the pacient	Fasting blood glucose	Tryglic-erydes	Hemoglobin type-A1c	Colesterol	Low density lipo-protein	Hight density lipo-protein
Age of the pacient	Pearson Correlation	1	-.020	-.038	.082	.052	.016	.085
	Sig. (2-tailed)		.779	.596	.251	.461	.821	.234
	N	200	200	200	200	200	200	200
Fasting blood glucose	Pearson Correlation	-.020	1	.264**	.820**	.199**	.188**	-.092
	Sig. (2-tailed)	.779		.000	.000	.005	.008	.197
	N	200	200	200	200	200	200	200
Tryglicerydes	Pearson Correlation	-.038	.264**	1	.340**	.354**	.226**	-.168*
	Sig. (2-tailed)	.596	.000		.000	.000	.001	.017
	N	200	200	200	200	200	200	200
Hemoglobin type-A1c	Pearson Correlation	.082	.820**	.340**	1	.187**	.151*	-.100
	Sig. (2-tailed)	.251	.000	.000		.008	.033	.158
	N	200	200	200	200	200	200	200
Cholesterol	Pearson Correlation	.052	.199**	.354**	.187**	1	.865**	-.043
	Sig. (2-tailed)	.461	.005	.000	.008		.000	.549
	N	200	200	200	200	200	200	200
Low density lipo-protein	Pearson Correlation	.016	.188**	.226**	.151*	.865**	1	-.034
	Sig. (2-tailed)	.821	.008	.001	.033	.000		.632
	N	200	200	200	200	200	200	200
Hight density lipo-protein	Pearson Correlation	.085	-.092	-.168*	-.100	-.043	-.034	1
	Sig. (2-tailed)	.234	.197	.017	.158	.549	.632	
	N	200	200	200	200	200	200	200

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 2.2 Correlation coefficients of independent variables

We see that the most important correlation is fasting blood glucose with hemoglobin type-A1c, low density lipo-protein with cholesterol. This means that the relationship between these two variables is significant.

Table 2.3 presents the results of the PCA method using (factory analysis) from SPSS, the first part of the table (Initial Eigenvalues) presents the numerical characteristics for all the selected components. While in the second part (Extraction Sums of Squared Loadings) the same numerical characteristics are presented, but only for those main components with a characteristic root greater than 1.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.473	35.333	35.333	2.473	35.333	35.333
2	1.489	21.271	56.604	1.489	21.271	56.604
3	1.110	15.856	72.460	1.110	15.856	72.460
4	.908	12.975	85.435			
5	.729	10.412	95.847			
6	.168	2.394	98.240			
7	.123	1.760	100.000			

Extraction Method: Principal Component Analysis.

Table 2.3 Results of the PCA method

The first component explains 35.333%, the second explains 21.271% and the third about 15.856%. The first component and the second component together explain 56.604% of the variance. While the three components together explain the variance 72.460%. The "Cumulative%" column expresses the accumulated percentages of the dispersion which is explained by the components where our model will be built on these three defined components.

To see the above conclusions in another, perhaps clearer way, figure 2.2 helps us by showing us a couple of points that correspond to the eigenvalues and the number of components as in the table.

This graph helps us to show the most important components as we see them and, in the graph, the first 3 components are further away than the other 2 components (most distinct) and the first 2 together with the third have the highest eigenvalue se 1. This graphic representation is used to locate the most important components.

Figure 2.2 Graphic representation of eigenvalues

Using the PCA method, we obtain Table 2.4, which shows the correlation coefficients between the initial variables and the principal components extracted by the PCA technique. The results of the PCA make it possible to group the initial variables according to the strength of the connection with the main components. We see that the first variable "cholesterol" has a correlation of 0.750 with the first component, while it has a smaller correlation with the other components, so we say that is related to the first component.

In this way, we group the other variables as follows:

- The variables "cholesterol", "hemoglobin type-A1c", "fasting blood glucose", "low density lipo-protein", "triglycerides" are related to the first component.
- The variable "age of the patient", "high density lipo-protein" is related to the third component.

	Component		
	1	2	3
Colesterol	.750	.607	.025
Hemoglobin type-A1c	.716	-.599	.182
Fasting blood glucose	.709	-.575	.131
Low density lipo-protein	.701	.633	.029
Tryglicerydes	.602	-.054	-.262
Age of the pacient	.038	.041	.734
High density lipo-protein	-.197	.160	.671

Table 2.4 Correlation of variables with components

Component Matrix.

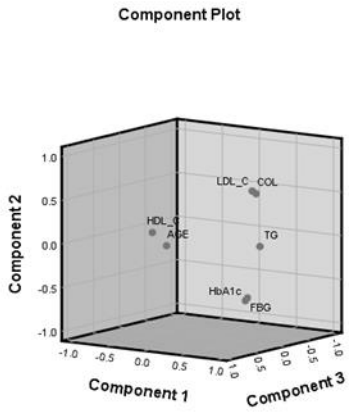

Figure 2.3 Graphical presentation of variables against 3 components

Figure 2.3 shows the relationship between the variables studied and the components. The points represent the variables with coordinates values of their correlations with the principal components, while the components represent the axes. What we are looking for are sets of variables that are centered around the axes.

As we can see, the first component has many variables, the third has 2 variables and the second none. The graph proves this. For this reason, we will use Factorial Analysis to regroup the factors.

CONCLUSION

- Men are the category most affected by heart attacks, spcifically coronary artery disease.
- Individuals who smoke are more predisposed to have narrowing of the blood vessels and therefore the presence of coronary artery disease.
- The age most affected by this type of disease is the average age, around 60.
- The most important factors in coronary diseases are: Hemoglobin sugar, Age and Gender. HbA1c is proposed as a reliable tool not only in the identification of diabetes but can also serve as a parameter for the risk of cardiovascular diseases.
- Our model will have the form: CAD = -0.514 + 0.106 *HbA1c + 0.017*Age - 0.377* Sex
- The smaller the value of sugared Hemoglobin and the age, the lower the CAD value will be for each of the sexes. It will be even smaller for women.
- Finding similarity in Chernoff faces is a bit difficult since the number of individuals is large.
- We see that PCA and FA give us the same conclusions, so their interpretation and results are more or less the same, but the mathematical functions with which we get these results are different. In both cases, the three components together explain the variance 72.460 %.
- From the 11 independent variables that we obtained through PCA and FA, we conclude that the best model is the one composed of 3 main components. So, the number of variables decreases from 11 to 3.

REFERENCES

[1] K. ARGJIRI, E. DHAMO (GJIKA), "Një model matematik mbi rastet e sëmundjeve kardiovaskulare në popullsinë e Shqipërisë", Buletini I Shkencave FSHN, 2012.
[2] A. John Camm, Thomas F. Lüscher, Gerald Maurer, Patrick W. Serruys," The ESC Textbook of Cardiovascular Medicine", Third Edition, OXFORD, 2018.
[3] P. Kohli, "Everything you need to know about heart disease", Medical News Today, 2021. Link: https://www.medicalnewstoday.com/articles/237191.
[4] M. Zagumny, The SPSS® Book: A Student Guide to the Statistical Package for the Social Sciences, Writers Club Press, 2001.
[5] S. Landau, B. S. Everitt "A handbook of statistical analyses using SPSS". Chapman & Hall/CRC Press LLC, 2004.
[6] P. Bruce and A. Bruce, "Practical Statistics for Data Scientists", O'Reilly Media, Inc. 2017.