International Journal of Engineering & Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-3444

International Journal of Engineering and Technology Volume 2 No. 8, August, 2012

A Comparative Study on Logistic Regression Model and PCA-Logistic Regression Model in Medical Diagnosis

J.Padmavathi
Computer Science, SRM University, Chennai, Tamil Nadu, 600 026, India

ABSTRACT

Logistic regression is a simple statistical tool used in binary classification. LR is useful for situations in which we want to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. Unlike LR, PCA is another technique used in feature extraction and in dimensionality reduction. This paper puts forth the experimental results of LR model and PCA-LR model in medical diagnosis, namely the CHD dataset.

Keywords: *Logistic regression(LR), PCA, feature extraction, Coronary Heart Disease(CHD).*

1. INTRODUCTION

Data Mining is an analytical process that is used to explore enormous data for pattern recognition or to identify the relationship between variables to generate new patterns. This predictive data mining is being increasingly used in business applications and in research [3,4]. The analytical process involves three stages, namely, exploration, Model building or pattern recognition and application of the model to new data in order to generate predictions.

Exploration is the initial stage that involves data preprocessing through data cleaning, transformation, subset or feature selection using statistical methods such as regression, factor analysis etc... The model building or pattern recognition involves designing the best model for predictive performance. In this stage different kinds of models are designed and evaluated for performance measures like accuracy, specificity and sensitivity. Data mining techniques such as bagging and boosting can be applied.

The last stage is application of the model to new data in order to generate predictions. This is to test the designed model and to estimate the expected outcomes. Logistic regression is used to analyze relationships between a dichotomous dependent variable or dichotomous independent variables. LR combines the independent variables to estimate the probability that a particular event will occur, i.e. a subject will be a member of one of the groups defined by the dichotomous data classes is as close as possible to the original distribution obtained using all attributes. Mining on the reduced set of attributes has additional benefits. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand. The variate or value produced by logistic regression is a probability value between 0.0 and 1.0. If the probability for group membership in the modeled category is above some cut point (the default is 0.50), the data is predicted to be a member of the modeled group, otherwise as member of another group. For any given case, logistic regression computes the probability that a case with a particular set of values for the independent variable; is a member of the modeled category.

$$Yi = e^{u}/(1 + e^{u})$$

Where Yi is the estimated probability that the i^{th} case is in a category and u is the regular linear regression equation:

$$u = A + B_1X_1 + B_2X_2 + \cdots + B_KX_K$$

2. RELATED WORK

The data about Coronary heart disease was collected and used in classifying the patients with heart attack was collected from Neomed Hospital. Attribute entries were recorded from Treadmill test, 3D-CCD and from physician examination recorded in the case history. Table 1, shows the attributes used for the analysis. The stepwise analysis used Exangi, Chol. Level, CP, slope (ST-depression or ST-elevation) as significant variables. Direct analysis used all the 17 parameters. SPSS software is used for analysis. Optimization criteria are taken as likelihood criteria. The single predictor is defined as,

$$p(x) = P(Y_j = 1 \; X = x) \Rightarrow E(Y_j X = x)$$

Instead of the probability of heart disease, we consider the odds as a function of age. Odds range from zero to infinity, so the problem fitting a linear model to the upper asymptote can be eliminated. If we go one step further and consider the logarithm of the odds, we now have a dependent variable that ranges from -1 to +1. We try to fit a linear regression model to the log-odds variable.

Our model would now be,

$$\text{logit}(p(x)) = \left(\log\left(\frac{p(x)}{1-p(x)}\right) = \beta 0 + \beta 1\right) \quad (1)$$

If we can successfully fit this linear model, then we also have successfully fit a nonlinear model for p(x), since the logit function is invertible, so after taking logit^{-1} of both sides, we obtain

$$p(x) = \text{logit}^{-1}(\beta_0 + \beta_1 x) \quad (2)$$

where,

$$\text{logit}^{-1}(w) = \exp(w)/(1 + \exp(w)) = 1/(1 + \exp(-w)) \quad (3)$$

The above system generalizes to more than one predictor, i.e,

$$p(x) = E(Y \mid X = x) \Rightarrow \text{logit}^{-1}(\beta' x) \quad (4)$$

It turns out that the system we have just described is a special case of what is now termed a generalized linear model. The Change in probability is not constant (linear) with constant changes in X. This means that the probability of a success (Y = 1) given the predictor variable (X) is a non-linear function, specifically a logistic function. The idea of comparing logistic regression with other methods is done by many researchers.[10]. This powerful method is used in classification process.

Table 1: Attributes of Cardiovascular disease dataset

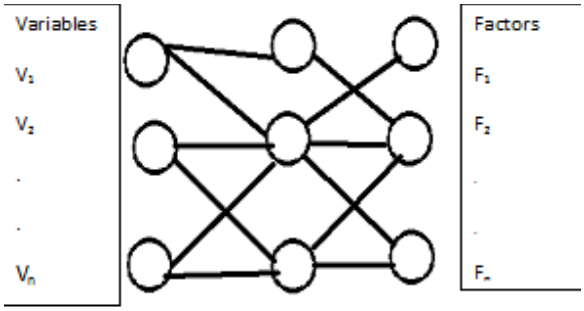
No	*Name*	*Description*
1	Age	Age in years
2	Sex	1 = male, 0 = female
3	Cpain	Chest pain type (1 = typical angina, 2 =atypical angina, 3 = non anginal pain,4 = asymptomatic)
4	RestBP	Resting blood Pressure (in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
7	RestECGg	Resting electrocardiographic results(0 = normal, 1 = having ST/ T wave abnormality, 2 = left ventricular hypertrophy)
8	MaxHR	Maximum heart rate
9	Exangi	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1 = up sloping, 2 = flat, 3 =down sloping)
12	Noves.	Number of major vessels colored by fluoroscopy
13	Thal	3 = normal, 6 = fixed defect, 7 = reversible defect
14	class	Class (0 = healthy, 1 = have heart disease)
15	Smoking	(yes/no)
16	Alcohol	(yes/no)
17	Diabetes	(yes/No)

DATA SET

Initial	Age	Sex	Chest PAIN TYP	Rest BP	Cho	FBS	Rest EC	Max HR	Exercis ANGINA	Old Pea up, flat,	The slo peak ex	No of v LAD,LCX,RCA	Thal	The Cla with heart disease	smokin	alco	Diabe
A.Mar	46	m	angina	110/80	190	104	T wave	154	TRUE	down	1.7 mm i	3		sick	yes	no	yes
S.Bal Mu	38	m	asympt	100/80	187	100	normal	142	TRUE	up	2.7 mm i	1 LAD		sick	yes	yes	no
k.gang	56	m	angina	130/80	141	129	ST eleva	148	TRUE	up	2.4 mm i	3		sick	no	no	yes
N.Rajas	55	m	angina	110/80	120	95	T wave	162	TRUE	down	1.9 mm i	1 lad		sick	no	no	no
j.swami	56	m	asympt	150/90	122	128	normal	156	TRUE	up	>2.0 mm	3		sick	no	no	no
c.arumu	69	m	angina	130/80	154	87	normal	146	TRUE	flat		2 lad/lcx		sick	no	no	no
Lil jaya	70	f	angina	110/70	230	95	T wave	152	TRUE	down	2.9 mv in	3	rever	sick	no	no	yes
S.M.Th p	63	m	asympt	130/90	88	171	st invers	149	TRUE	down	2.8 mm i	3	rever	sick	no	no	no
a.arunar	65	f	angina	130/90	138	92	st invers	145	TRUE	down	1.95 mv	3		sick	no	no	no
j.manmul	62	m	angina	150/90	194	224	normal	160	TRUE	flat		3		sick	no	no	no
gom ram	56	f	not ang	110/80	212	200	normal	158	FALSE					buff	no	no	yes
s.p.pillap	59	m	angina	120/90	103	164	normal	164	TRUE	up	2.6 mm a	3		sick	no	no	no
g.sridha	46	m	angina	120/80	187	89	st invers	157	TRUE	down	2.9 mv in	3		sick	yes	yes	yes
g.k.srini	72	m	asympt	160/90	127	170	normal	140	TRUE	flat		3		sick	no	no	no
p.bal	62	m	angina	140/90	125	146	normal	165	TRUE	up	2.9 mm i	3		sick	no	no	yes
v.ramkri	72	m	angina	120/90	177	238	ST eleva	142	TRUE	up	>2.0 mm	3		sick	no	no	yes
v.kalper	59	m	angina	120/90	151	82	ST eleva	161	TRUE	up	>2.0 mm	2 lad,rca		sick	no	no	yes
k.govinra	65	m	angina	160/90	173	94	normal	153	TRUE	down	<2.0 mv	3	rever	sick	no	no	no
s.subr re	66	m	angina	110/80	166	170	normal	164	TRUE	flat		3		sick	no	no	no
s.xavie	65	m	angina	140/80	137	90	st invers	155	TRUE	down	2.95 mv	2 lad,lcx		sick	no	yes	yes
s.nar rad	56	m	angina	120/80	160	124	normal	168	TRUE	up	>2.0 mm	2 lad,rca		sick	no	no	yes
S.A.ISM/	87	M	angina	160/80	176	90	st invers	138	TRUE	down	2.6 mv in	3		sick	yes	no	no

2.1 PCA-LR Model

Principal Component Analysis (PCA) networks are a mixture of unsupervised and supervised networks The unsupervised segment of the network performs the feature extraction and the supervised segment of the network performs the (linear or nonlinear) classification of these features



PCA Network

The projection variable y isgiven as, $y = W^T x$, where W is nxm matrix that contains the principal components as columns. The vector $\mathbf{y}$ is a dimension-reduced representation of $\mathbf{x}$. Let $\hat{\mathbf{x}}$ be the reconstruction of $\mathbf{x}$ given only the vector $\mathbf{y}$, $\hat{\mathbf{x}} = W y$.

The covariance matrix $\mathbf{C}$ of the pattern set is computed as, $\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

The eigen values are computed as, $\mathbf{C}\,\mathbf{w_l} = \lambda_l\,\mathbf{w_l}$

The original 'n' features of the data set are replaced by 'm' new features of data that are formed from linear combinations of the original features. This model is used to determine performance measure, transformation and a number of new features. To form a 'm'-dimensional projection 1<m<n-1 by those linear combinations that maximize the sample variance subject to being uncorrelated with all these already selected linear combination. . The projections of these components correspond to the eigenvalues of the input covariance matrix. The aim is to capture intrinsic variability in data through linear mapping from original features to newly extracted ones. The number of new features 'm' principal components that capture the amount of variance subject to predetermined threshold; the transformation is linear combination. The performance measure is sample variance. PCA does not require the data to be labeled with classes. The benefit of this dimensionality reduction includes providing a simpler representation of data reduction in memory and faster classification[1,2].

PCA-LR model works in two stages. The PCA model is used to extract principal components and is fed into LR model for classification. Out of 16 variables used for the study. PCA network extracted 6 components for analysis. These values were fed into Logistic regression and the results are summarized. The process was repeated for Two sets of data, namely 257 and 177 records.

3. CONCLUSION

Logistic regression is a type of multivariable analysis used with increasing frequency in the health sciences because of its ability to model dichotomous outcomes. Proper use of this powerful and sophisticated modeling technique requires considerable care both in the specification of the form of the model and in the calculation and interpretation of the model's coefficients[13][14]. The coefficients of the predictor variables are interpreted as signifying the relative contribution of their respective variables toward the predicted probability of a positive outcome. A three-way comparison of prediction accuracy involving nonlinear regression, NNs and CART models using a continuous dependent variable and a set of dichotomous and categorical predictor variables was studied by many researchers[11,12]. This is slightly varied by identifying the number of parameters that are to be included in the prediction process.

The analysis shows that classification using LR was better than PCA-LR. This is because negative values were more in number than the positive. The criteria considered in this article can affect the regression coefficients, in different ways and at different stages of the model-building process. Although many parts of the process have been effectively automated, the authority of the final model depends on the attempts by investigators to rule out sources of bias or inaccuracy toward which each of the criteria contributes [15][16]. Medical data mining can help to prepare some methods for diagnosis, prognosis, decision making, etc[17]. The results of PCA-LR networks are appreciable only if feature extraction is good. The aim of the study was to design an intelligent computer based neural network model that will enable successful prediction of the heart disease.

ACKNOWLEDGEMENT

The author expresses her gratitude to Dr. Sivabalan, Director & chief Consultant Preventive Cardiology, Chelation Therapy NEOMED Hospital, Member - American College of Advanced medicine (ACAM) and Dr. Tirumavalan of Neomed hospital, Speciality and Research Center for their support

in the research by providing data and Dr. R.Chandrasekar, Head Dept. Of Statistics for his constant support throughout the study.

Classification result using logistic regression

Data size 257 records. The number of records with absence was 143, while that of presence was 114.

Classification Table(a)

Observed			Predicted		
			class		Percentage Correct
			1.00	2.00	
Step 1	class	1.00	129	14	90.2
		2.00	17	97	85.1
	Overall Percentage				87.9

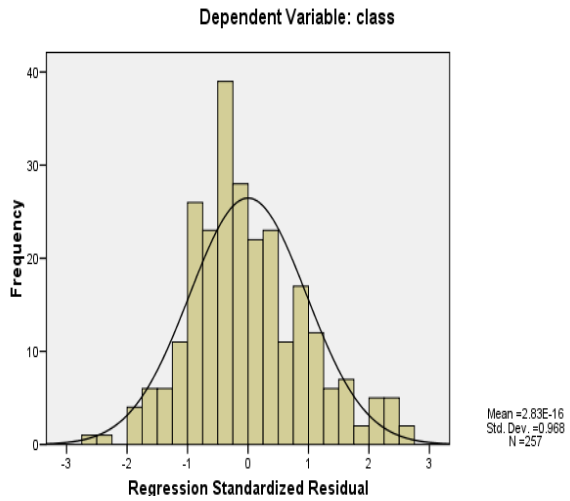
a The cut value is .500

Variables in the Equation

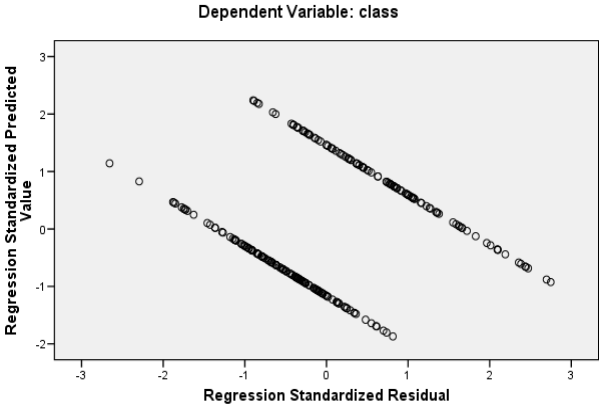
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	age	.000	.029	.000	1	.997	1.000
	sex	1.613	.590	7.469	1	.006	5.019
	cpain	.914	.254	12.916	1	.000	2.495
	restbp	.028	.012	5.213	1	.022	1.028
	chol	.011	.005	4.308	1	.038	1.011
	fbs	15.862	40192.948	.000	1	1.000	7738929.269
	restecg	.335	.223	2.255	1	.133	1.398
	maxhr	-.028	.012	5.356	1	.021	.972
	exerangina	.830	.481	2.978	1	.084	2.294
	oldpeak	.634	.439	2.083	1	.149	1.885
	slope	.088	.253	.121	1	.728	1.092
	noves	1.127	.295	14.615	1	.000	3.086
	thal	.874	.242	13.045	1	.000	2.398
	smoke	.472	.855	.306	1	.580	1.604
	alco	1.956	1.115	3.078	1	.079	7.070
	diab	-16.806	40192.948	.000	1	1.000	.000
	Constant	-10.386	3.533	8.643	1	.003	.000

a Variable(s) entered on step 1: age, sex, cpain, restbp, chol, fbs, restecg, maxhr, exerangina, oldpeak, slope, noves, thal, smoke, alco, diab.

Histogram



Scatterplot

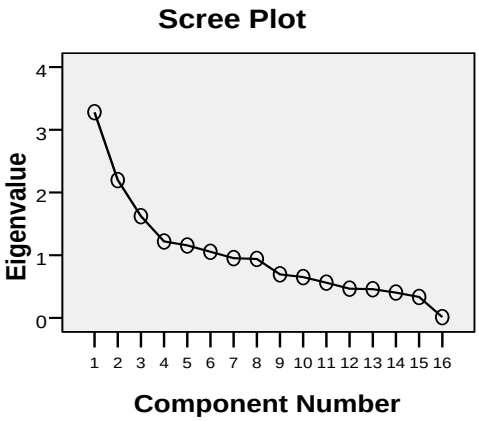


RESULTS OF PCA_LR

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
age	54.1440	9.10993	257
sex	.6732	.46998	257
cpain	3.1712	.94882	257
restbp	131.4202	18.12938	257
chol	248.0545	47.80900	257
fbs	.1829	.38732	257
restecg	1.0117	.99798	257
maxhr	149.9650	23.22854	257
exerangina	.3268	.46998	257
oldpeak	1.5837	.62014	257
slope	1.0533	1.14598	257
noves	.6615	.94279	257
thal	1.7899	.95331	257
smoke	.0973	.29691	257
alco	.0739	.26217	257
diab	.1790	.38409	257

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.280	20.498	20.498	3.280	20.498	20.498	2.074	12.960	12.960
2	2.198	13.737	34.235	2.198	13.737	34.235	2.031	12.691	25.651
3	1.621	10.133	44.368	1.621	10.133	44.368	1.933	12.081	37.732
4	1.219	7.619	51.987	1.219	7.619	51.987	1.553	9.704	47.437
5	1.155	7.221	59.208	1.155	7.221	59.208	1.492	9.328	56.764
6	1.054	6.589	65.797	1.054	6.589	65.797	1.445	9.033	65.797
7	.952	5.952	71.750						
8	.940	5.875	77.625						
9	.694	4.339	81.964						
10	.651	4.069	86.033						
11	.561	3.506	89.539						
12	.466	2.912	92.451						
13	.458	2.862	95.313						
14	.404	2.525	97.838						



Component Matrix[a]

	Component					
	1	2	3	4	5	6
age	.402	.420	-.433	-.016	.105	-.156
sex	.295	-.150	.533	-.196	-.121	.513
cpain	.482	-.115	-.148	-.530	-.025	-.171
restbp	.239	.359	-.296	.389	.118	.367
chol	.112	.211	-.492	-.141	.495	.218
fbs	-.068	.884	.408	-.013	.057	-.099
restecg	.229	.091	-.250	.181	.141	.456
maxhr	-.673	-.054	.183	.101	.190	.339
exerangina	.615	-.069	.024	-.232	-.032	.083
oldpeak	.622	.050	-.009	.466	-.362	-.123
slope	.720	-.005	.023	.398	-.191	-.057
noves	.464	.261	-.137	-.404	.058	.014
thal	.596	.042	.316	-.197	-.148	.335
smoke	.416	-.227	.350	.145	.573	-.257
alco	.428	-.304	.378	.176	.536	-.099
diab	-.092	.894	.386	-.033	.040	-.096

Extraction Method: Principal Component Analysis.
a. 6 components extracted.

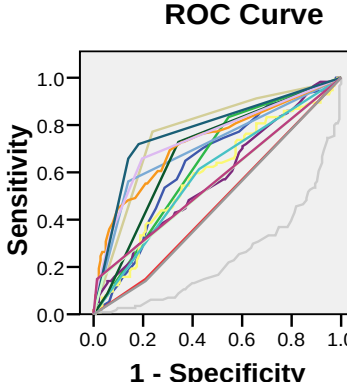
Case Processing Summary

class(b)	Valid N (listwise)
Positive(a)	114
Negative	143

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

a. The positive actual state is 2.00.

b. The test result variable(s): maxhr has at least one tie between the positive actual state group and the negative actual state group.



Diagonal segments are produced by ties.

Logistic applied on extracted parameters

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	257	100.0
	Missing Cases	0	.0
	Total	257	100.0
Unselected Cases		0	.0
Total		257	100.0

a. If weight is in effect, see classification table for the total number of cases.

Classification Table(a)

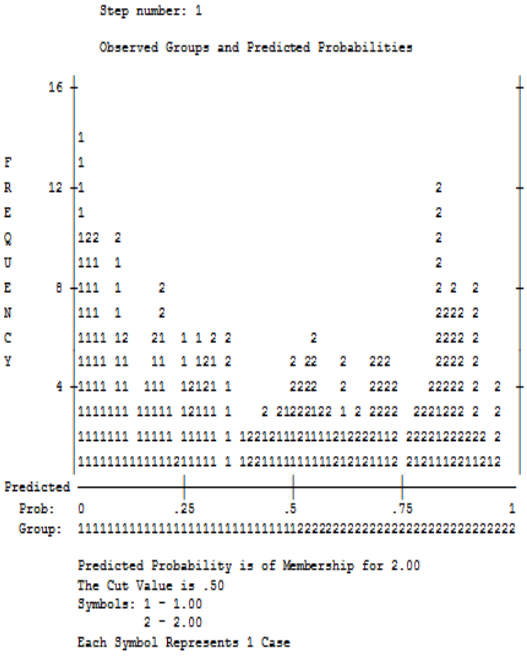
Observed			Predicted		
			class		Percentage Correct
			1.00	2.00	
Step 1	class	1.00	115	28	80.4
		2.00	25	89	78.1
	Overall Percentage				79.4

a The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age	.062	.021	9.124	1	.003	1.064
	sex	2.294	.414	30.653	1	.000	9.915
	cpain	1.387	.213	41.299	1	.000	3.922
	restbp	.024	.009	6.809	1	.009	1.024
	chol	.007	.003	3.850	1	.050	1.007
	fbs	-.554	.425	1.703	1	.192	.575
	Constant	-14.512	2.125	46.628	1	.000	.000

a. Variable(s) entered on step 1: age, sex, cpain, restbp, chol, fbs.



DATA SET II

The number of records were 177. The number of positives were 170 and negatives were 7.

The classification results are as follows. With Logistic regression the classification percentage waS 100.

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	177	100.0
	Missing Cases	0	.0
	Total	177	100.0
Unselected Cases		0	.0
Total		177	100.0

a If weight is in effect, see classification table for the total number of cases

Classification Table(a)

Observed			Predicted		
			class		Percentage Correct
			1.00	2.00	
Step 1	class	1.00	7	0	100.0
		2.00	0	170	100.0
	Overall Percentage				100.0

a The cut value is .500

With PCA_LR the classification results with 5 parameters and with 12 parametrs are shown.

Classification Table(a)

Observed			Predicted		
			class		Percentage Correct
			1.00	2.00	
Step 1	class	1.00	0	7	.0
		2.00	0	170	100.0
	Overall Percentage				96.0

a The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	age	.033	.048	.466	1	.495	1.033
	sex	1.215	.833	2.129	1	.145	3.370
	cpain	-.357	.330	1.173	1	.279	.699
	RestBP	-.012	.023	.255	1	.613	.988
	chol	.004	.007	.284	1	.594	1.004
	Constant	2.137	3.581	.356	1	.551	8.472

a Variable(s) entered on step 1: age, sex, cpain, RestBP, chol.

The PCA extracted features namely age, sex, chest pain type, BP at rest and cholesterol level are taken for prediction. Percentage of classification was 96.

Observed			Predicted		
			class		Percentage Correct
			1.00	2.00	
Step 1	class	1.00	3	4	42.9
		2.00	2	168	98.8
	Overall Percentage				96.6

a The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	age	-.123	.104	1.409	1	.235	.884
	sex	1.851	1.838	1.014	1	.314	6.366
	cpain	.998	.786	1.611	1	.204	2.712
	RestBP	.003	.032	.007	1	.935	1.003
	chol	.029	.017	2.913	1	.088	1.029
	FBS	-.362	1.229	.087	1	.768	.696
	RestECG	-1.044	1.416	.543	1	.461	.352
	MaxHR	.052	.042	1.568	1	.210	1.054
	exerangi	3.432	1.779	3.722	1	.054	30.953
	oldpeak	-.840	.943	.794	1	.373	.432
	slope	1.135	.943	1.451	1	.228	3.112
	noves	3.081	1.477	4.349	1	.037	21.779
	Constant	-10.563	9.477	1.242	1	.265	.000

a Variable(s) entered on step 1: age, sex, cpain, RestBP, chol, FBS, RestECG, MaxHR, exerangi, oldpeak, slope, noves.

Future Work

The future work is to design a hybrid model, which would help in predicting heart attack. Feature extraction is done with PCA network and the data is fed into RBF neural network.

REFERENCES

[1] Vollmer RT. Multivariate statistical analysis for pathologists. Part I, The logistic model. Am J Clin Pathol 1996;105:115–26.

[2] Patterson DW. Artificial neural networks. New York: Prentice Hall; 1995

[3] E., Donald, "Introduction to Data Mining for Medical Informatics ," *Clin Lab Med*, pp. 9-35, 2008.

[4] R., Zhang, Y., Katta, "Medical Data Mining," *Data Mining and Knowledge Discovery*, pp. 305-308, 2002.

[5] Irene M. Mullins et al., "Data mining and clinical data repositories: Insights from a 667,000 patient data set," *Computers in Biology and Medicine*, vol. 36, pp. 1351-1377, 2006.

[6] J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., EdwardHammond, W. Parther, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse," 1997.

[7] Lemeshow S, Hosmer DW. Logistic regression. In: Armitage P, ColtonT, Eds. Encyclopedia of Biostatistics. New York: J. Wiley, 1998.p. 2316–27.

[8] Glantz SA, Slinker BK. Primer of applied regression and analysis of variance. New York: McGraw-Hill, Inc., 1990.

[9] Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley, 1989

[10] Abu-Hanna, A., & de Keizer, N. (2003). Integrating Classification trees with local logistic regression in intensive care prognosis. Artificial Intelligence in Medicine, 29, 5-23.

[11] Khemphila, A.; Boonjing, V., "Comparing the performanceof logistic regression, decision trees and neural networks for classifying heart disease patients". Proceedings of International Conference on Computer Information System and Industrial Management Applications 2010, pp. 193 –198.

[12] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[13] Kurt, I.; Ture, M.; Turhan, A., "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". Journal of Expert Systems with Application, Vol3,2008,pp.366-374.

[14] Linder R, Konig IR, Weimar C, Diener HC, Poppl SJ, Ziegler A. Two models for outcome prediction – a comparison of logistic regression and neural networks. Methods Inf Med 2006;45(5):536–40.

[15] Harrel Jr FE. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. Berlin: Springer; 2001.

[16] Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. Washington, DC: CRC Press; 1997.

[17] Ljiljana Trtica-Majnaric , Marijana Zekic-Susac , Natasa Sarlija , Branko Vitale Prediction of influenza vaccination outcome by neural networks and logistic regression Journal of Biomedical Informatics 43 (2010) 774–781