A Linear Model based on Principal Component Analysis for Disease Prediction

Shreerekha M[1], Mrs. Padmanayana[2]

[1]Computer science and Engineering Srinivas Institute of Technology, Valachil (NAAC Accredited) Mangaluru, Karnataka, India
Associate Prof. Dept. of Computer Science and Engineering
[2]Srinivas Institute of Technology, Valachil (NAAC Accredited) Mangaluru, Karnataka, India

***

Abstract - *Analysis of a disease is tough task in the medical field. Various classification methods are used to predict the diseases. Diagnosis of diabetes can be analyzed by checking the level of blood sugar of patient with the normal known levels, blood pressure, BMI, skin thickness, and so on. The main aim of this paper is to build a statistical model to predict the diabetes. The feature extracted using Principal Component Analysis and then modeled using Linear Regression Model. The accuracy obtained by this method is 82.1 % for predicting diabetes.*

***Key Words*: Principal Component Analysis, Linear Regression Model, Diabetes, Pima Indian Diabetes Data.**

1.INTRODUCTION

Analysis of diseases is a difficult task in medical field. **Diabetes** is a metabolic disease that causes high blood sugar.The paper presesnts the stastistical model to predict the diabetes. Principal Component Analysis(PCA) is used to extract feature and then modeled using Linear Regression Model.The dataset used is Pima Indian diabetes dataset(PIDD).

Feature extraction is main step in examining the PIDD dataset. PCA is reduction method which considers the PIDD as set of rows representing characteristics in a high dimensional space and all rows are put up to a directions which represents the best set of features.The original features of PIDD are approximated with fewer dimensions which are an overall of original PIDD using PCA.The model is build using linear regression model.

2.RELATED WORKS

H. Roopa, T. Asha[1] proposed a A Linear Model Based on Principal Component Analysis for Disease Prediction. It uses PCA and LRM.Polat et al. [2] proposed a Least Square Support Vector Machine (LS-SVM) classification method to obtain an accuracy of 79.16%. Generalized Discriminant Analysis (GDA) was used at preprocessing stage for discriminating variables of PIDD and then LS-SVM technique was applied.

Seera and Lim [3] proposed Fuzzy Min–Max neural network, Regression Tree and Random Forest (FMM-CART-RF) combined hybrid classification method that achieved an accuracy of 78.39% for PIDD.

Sa'di et al. [4] classified PIDD using various data mining algorithms and analyzed that Naïve Bayes performed well than RBF network and J48 with an accuracy of 76.95%.Bansal et al. [5] proposed an evolutionary method where feature selection of PIDD is obtained by Particle Swarm Optimization (PSO) method and then k-Nearest Neighbor (KNN) classification technique is applied on these features to achieve an accuracy of 77%. Choubey et al. [6] applied Genetic Algorithm (GA) for variable selection on PIDD. The accuracy of 78.69%.

3.PROPOSED METHODOLOGY

The proposed methodology includes feature extraction of PIDD using PCA and then modeling using LRM. The model illustrated using the Figure1.

A. Input data

The proposed system make use of Pima Indian Diabetes Data. PIDD consist of 8 actual variables and one class variable. There are total 768 instances out of which 268 instances have class variable value '1' and 500 instances have class variable value '0' respectively. If the binary response variable represents '1' means ''positive for diabetes'' and '0' means ''negative for diabetes''. The attributes of PIDD are presented in Table 1.

B. FEATURE EXTRACTION BY PCA

Principal Component Analysis is applied to extract the feature of PIDD values to a new space. The following steps are involved:

- Subtract the mean from each dimensions of PIDD. It will produces a data set whose mean is zero.
- Calculate the variance matrix between two separate dimensions of PIDD.
- Find the Eigen vector and Eigen values of the matrix obtained in step2.
- Construct the feature vector and take transpose of it. Then multiply it with original PIDD to obtain new set of features projected to new space.

Table -1: The attributes of PIDD

No.	Variables	Values
1	Pregnancies	Integer
2	Glucose	Integer
3	Blood Pressure	Integer
4	Skin thickness	Integer
5	Insulin	Integer
6	BMI	Numeric
7	Diabetes Function	Numeric
8	Age	Integer
9	Class	Integer(0 or 1)

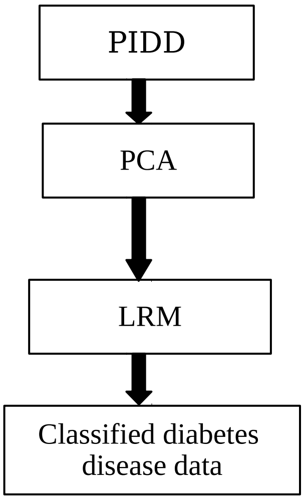


Figure1. Steps involved in the proposed system.

C. LINEAR REGRESSION MODEL (LRM)

The regression analysis has dependent variable value ‘y’ and independent variable values‘$x_1,x_2,x_3,..,x_k$’.Linear regression model is represented by equation (1)

$$y=b_0+b_1x_1+b_2x_2+...+b_kx_k+\varepsilon \qquad (1)$$

where ‘y’ is the class variable of PIDD data,

b_0 is a ‘y’ intercept,

ε is an error term and

b_1 ,b_2 ,b_3,.....,b_k are coefficients of $x_1,x_2,x_3,...,x_k$. respectively.

‘y’ is based on the set of independent variable values $x_1,x_2,x_3,...,x_k$.

The fitted line of equation (1) is calculated by principle of least square method where b_1 ,b_2 ,b_3,.....,b_k are chosen in such a way that the Sum of Squares of Error (SSE) is minimum.

Consider equation (1), For n = 768 observation,

Let,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{738} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11}x_{21} \cdots & x_{81} \\ \vdots & \ddots & \vdots \\ 1 & x_{1786} \cdots & x_{8768} \end{bmatrix}$$

$$\hat{b} = \begin{bmatrix} \hat{b_0} \\ \hat{b_1} \\ \hat{b_2} \\ \vdots \\ \vdots \\ \hat{b_k} \end{bmatrix}$$

$\hat{b}$ → least square estimates of b_1 ,b_2 ,b_3,.....,b_k of linear model.

Least square matrix is obtained by

$(x^1x)\ \hat{b} = x^1y$

where x^1 → is transpose of x matrix.

(x^1x) → coefficient matrix of least square estimates of b_0 , $\hat{b}_1$ . . . . $\hat{b}_k$.

x^1y → gives matrix of constants.

Therefore least square solution is obtained by $\hat{b} =(x^1x)^{-1}.x^1y$

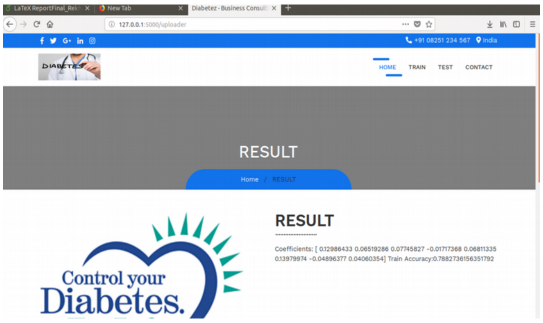
Substitute $\hat{b}$ in equation (1) to get the final fitted line.

D. OUTPUT DATA

The model classifies features of PIDD as diabetic or normal.

4.RESULTS

The attributes of PIDD is projected to a new space using PCA. Then LRM is applied on the components of PIDD. The results are shown below. The PIDD is partitioned such a way that 80% are used as training dataset and 20% are used for testing. That means 614 for training and 154 for testing. The confusion metrics for test data and ROC curve for training and testing data are shown below.



RESULT

Coefficients: [0.12986433 0.06519286 0.07745827 -0.01717368 0.06811335 0.13979974 -0.04896377 0.04060354] Train Accuracy:0.7882736156351792

Figure2. Result on webpage for train data

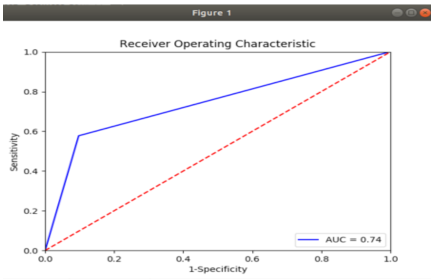
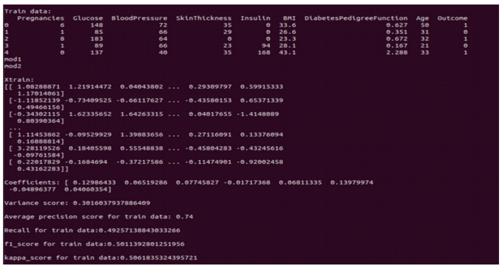
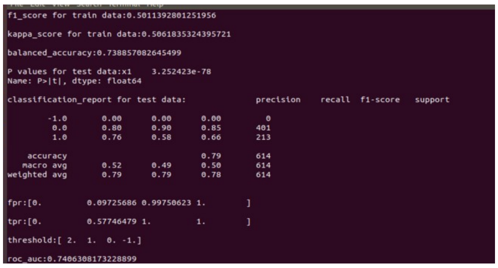


Figure3. ROC curve for train data



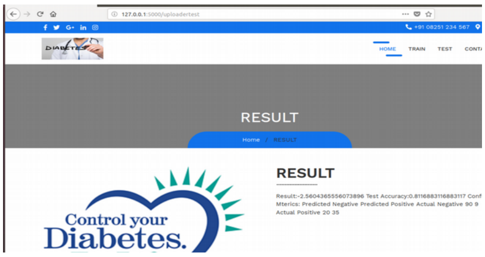
```
Train data:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
0            6      148             72             35        0  33.6                     0.627   50        1
1            1       85             66             29        0  26.6                     0.351   31        0
2            8      183             64              0        0  23.3                     0.672   32        1
3            1       89             66             23       94  28.1                     0.167   21        0
4            0      137             40             35      168  43.1                     2.288   33        1
mod1
mod2

Xtrain:
[[ 1.08288871  1.21914472  0.04043802 ...  0.29309797  0.59915333
   1.17014061]
 [-1.11852139 -0.73409525 -0.66117627 ... -0.43580153  0.65371339
   0.49466156]
 [-0.34302115  1.62335652  1.64263315 ...  0.04017655 -1.4148089
   0.80390364]
 ...
 [ 1.11453862 -0.09529929  1.39883656 ...  0.27116091  0.13376094
   0.16888814]
 [ 3.28119526  0.18405598  0.55548838 ... -0.45804283 -0.43245616
  -0.09761584]
 [ 0.22017829 -0.1684694  -0.37217586 ... -0.11474901 -0.92002458
   0.43162283]]

Coefficients: [ 0.12986433  0.06519286  0.07745827 -0.01717368  0.06811335  0.13979974
 -0.04896377  0.04060354]

Variance score: 0.3016037937886409

Average precision score for train data: 0.74

Recall for train data:0.49257138843033266

f1_score for train data:0.5011392801251956

kappa_score for train data:0.5061835324395721
```

Figure4. Result of train data.



```
f1_score for train data:0.5011392801251956

kappa_score for train data:0.5061835324395721

balanced_accuracy:0.738857082645499

P values for test data:x1    3.252423e-78
Name: P>|t|, dtype: float64

classification_report for test data:              precision    recall  f1-score   support

        -1.0       0.00      0.00      0.00         0
         0.0       0.80      0.90      0.85       401
         1.0       0.76      0.58      0.66       213

    accuracy                           0.79       614
   macro avg       0.52      0.49      0.50       614
weighted avg       0.79      0.79      0.78       614


fpr:[0.         0.09725686 0.99750623 1.        ]

tpr:[0.         0.57746479 1.         1.        ]

threshold:[ 2.  1.  0. -1.]

roc_auc:0.7406308173228899
```

Figure5. Result for classification report.



RESULT

Result:-2.5604365556073896 Test Accuracy:0.8116883116883117 Conf Mterics: Predicted Negative Predicted Positive Actual Negative 90 9 Actual Positive 20 35

Figure6. Result of test data on webpage

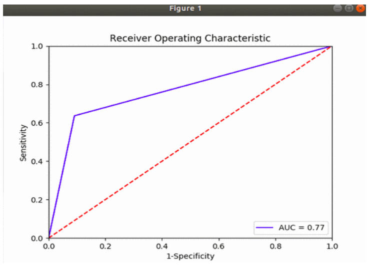
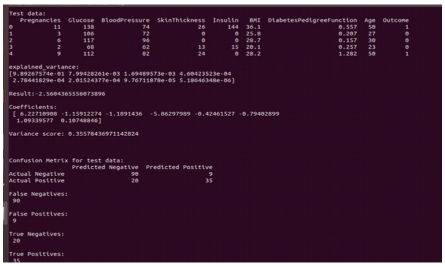


Figure7. ROC curve for test data



```
Test data:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
0           11      138             74             26      144  36.1                     0.557   50        1
1            3      106             72              0        0  25.8                     0.207   27        0
2            6      117             96              0        0  28.7                     0.157   30        0
3            2       68             62             13       15  20.1                     0.257   23        0
4            9      112             82             24        0  28.2                     1.282   50        1

explained_variance:
[9.89267574e-01 7.99428261e-03 1.69489573e-03 4.60423523e-04
 2.78441829e-04 2.01524377e-04 9.76711878e-05 5.18646348e-06]

Result:-2.5604365556073896

Coefficients:
 [ 6.22710908 -1.15912274 -1.1891436  -5.86297989 -0.42461527 -0.79402899
  1.09339577  0.10748846]

Variance score: 0.35578436971142824



Confusion Metrix for test data:
                 Predicted Negative  Predicted Positive
Actual Negative                  90                   9
Actual Positive                  20                  35

False Negatives:
 90

False Positives:
 9

True Negatives:
 20

True Positives:
 35
```

Figure8. Confusion metrics for test data

5.Conclusion and future work

This paper presents the feature extraction and statistical modeling on PIDD dataset. The PIDD data is extracted to new space using PCA. The newly projected features then modeled using LRM to predict whether the patient is diabetic or normal. The results obtained in this study have achieved high accuracy rate for predicting diabetes when compared with other existing methods. As future scope,the proposed statistical model can be adopted for predicting different kinds of diseases like tuberculosis, eye disease, cancers, etc.,

REFERENCES

[1] H. Roopa and T. Asha,'' A Linear Model Based on Principal Component Analysis for Disease Prediction',' Dig.Obj. Identifier10.1109/ACCESS.2019.2931956

[2] K. Polat, S. Güneş, and A. Arslan, ''A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine,'' Expert Syst. Appl., vol. 34, no. 1, pp. 482–487, 2008.

[3] M. F. Ganji and M. S. Abadeh, ''A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis,'' Expert Syst. Appl., vol. 38, no. 12, pp. 14650–14659, 2011.

[4] M. Seera and C. P. Lim, ''A hybrid intelligent system for medical data classification,'' Expert Syst. Appl., vol. 41, no. 5, pp. 2239–2249, 2014.

[5] S. Sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi, ''Comparison of data mining algorithms in the diagnosis of type II diabetes,'' Int. J.Comput. Sci. Appl., vol. 5, no. 5, pp. 1–12, 2015.

[6] R. Bansal, S. Kumar, and A. Mahajan, ''Diagnosis of diabetes mellitus using PSO and KNN classifier,'' in Proc. Int. Conf. Comput. Commun.Technol. Smart Nation (IC3TSN), Oct. 2017, pp. 32–38.