International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

# Prediction of Diabetics using Factor Analysis

**S. Kavitha, E. Srividhya, S. Muthuselvan**

***Abstract*: A diabetic is a fast growing disease in the world so prediction of diabetics is so important for quick decision making. The data mining techniques are used for analysis of medical database. The one of the data mining technique is statistical methods which is playing a major role for analysis and prediction of diabetics in accurate manner. The factor analysis is a method of reducing huge variables into lesser number of factors. It extracts the maximum common variances from all the variables and puts into the common variables. These common variables are used for further analysis. The factor analysis of dataset will give an effective outcome or better result to predict and also diagnose the diabetes disease. This paper focused on increasing the quality and accuracy of knowledge for diabetes disease treatment.***

***Index Terms*: Data Mining, Factor Analysis, Diabetic, prediction.***

## I. INTRODUCTION

The data mining is the process or method of knowledge discovery from databases. It refers to acquire the knowledge as well as the patterns from the large dataset. It gives the estimation and evaluation of patterns for quick decision making. It is producing the important results for researchers. It is an analytical process and also designed to explore the data in search of patterns as well as to find the systematic relationships among the variables.

The Diabetes mellitus is a group of the metabolic diseases of person who is having high blood sugar / glucose. It causes the insulin hormone deficiency in human body. So, it is called insulin deficiency. The diabetes is that pancreas is not producing enough insulin and the body cells are not responding to the produced- insulin in proper. Now-a-days the diabetics' patients are growing large amount in world wide. The diabetics' mellitus leads to various complications such as heart attract, eye damage, stroke, Kidney failure, Nerves disease etc. It needs to predict and develop the diabetes diagnosis system for society betterment and also helpful for the medical professionals for accurate decision making as well as the timely treatment to the patients.

The fields of medical diagnosis have huge volume of data generated. For solving the above problems, the data mining tools are used for analyzing of data for knowledge extraction. The data mining is an important role for taking decision in correct and accurate manner of healthcare industry. The medical data mining is used to find the useful pattern for medical diagnosis. It will help to predict the disease in earlier for patient care. The data pre-processing is required for preparation of data for mining, then reduce the amount of data to be analyzed without losing and also improves the data quality.

**Revised Manuscript Received on 30 March 2019.**
* Correspondence Author
**S. Kavitha***, Dept. of Computer Science and Technology, SRM Institute of Science and Technology, Chennai, India.
**E. Srividhya**, Dept. of Computer Science and Engineering, Aarupadai Veedu Institute of Technology, Chennai, India.
**S. Muthuselvan**, Dept. of Computer Science and Engineering, Aarupadai Veedu Institute of Technology, Chennai, India.



## II. LITERATURE REVIEW

### A. Data Mining

The potential of the medical data mining is to explore the hidden patterns in medical datasets. It is used for the clinical diagnosis. The decision support systems and appropriate computer-based information are used to achieve the clinical tests at the cost reduction basis as well as generate the efficient / accurate results for diagnosis. There are various methods / techniques are available for analyzing the datasets in data mining.

The K-NN, Decision list algorithm and Naïve Bayes of supervised machine learning algorithm are used to predict the disease. The Bayesian classification is for reducing the actual size of data and also get optimal subset of attribute for heart disease prediction [1].

Fuzzy Logic Techniques, Data Clustered Algorithms, Neural-Network Algorithm, and Hybrid Genetic Algorithm are applied in the proposed model for identifying diabetes mellitus, type of diabetes and its complications. It will reduce the cost for the variety of tests and provide the preventive measures well in advance [2].

The Naïve Bayes Classifier, Support Vector Machine, Principal Component Analysis algorithms and Decision Tree Algorithm are applied for the prediction of disease. The maximum ROC area means the excellent predictions performance [3].

The Hadoop MapReduce based Machine Learning Algorithms are for analysis of diabetes disease dataset for finding the missing values and also discover patterns from dataset [4].

### B. Factor Analysis

The factor analysis is a mathematical model and to estimate the relationship between observed indicators and latent variable by determining covariance or correlation among observable indicators [5]. Each variable is expressed as the linear combination of components or factors. The following is the model of factor analysis in the factor analysis.

$$X_i = B_{i1}F_1 + B_{i2}F_2 + B_{i3}F_3 + \ldots + B_{in}F_n + V_iU_i$$

Here,

$X_i - i^{th}$ standardized variable

$B_{ij}$ – Standardized multi-regression coefficient of the variable $i$ on the $j$

$F_j$ - Common factor $j$

$V_i$ - Standardized regression coefficient of variable $i$ on unique factor $j$

$U_i$ - The Unique factor for variable $i$

$n$ - Number of common factors

### C. Procedures:

| S. No | Procedures | Descriptions |
|---|---|---|
| 1. | Test of Bartlett (sphericity) | It is used for testing the hypothesis that variables are uncorrelated. |
| 2. | Matrix of Correlation | The lower triangle in the matrix is shown as correlations (*r*) among all the possible pair of the variables. |
| 3. | Communality | Amount of variable variance is that shares with all the other variables. |
| 4. | Eigen value | It represented total variance. |
| 5. | Loading Of Factors | It is loading the factors values and also shows the correlations among variables and factors. |
| 6. | Factor Matrix | It describes the loading of factors of all variables. |
| 7. | Scores of Factor | It is a composite-score and also estimated each respondent in the derived factors. |
| 8. | KMO - sampling adequacy measurement | It examines the appropriateness of the analysis of factor. The high values are from 0.5 to 1.0. |
| 9. | Variance Percentage | The total variance percentage is attributed to each of the factors. |
| 10. | Screeplot | Plot the Eigen values in the graph against the number of components in order to extract. |

Table 1 Procedures for Factor Analysis

## III. RESEARCH METHODOLOGY

The dataset is taken from database of the Prima Indian Diabetes from UCI Machine Learning Repository. It consists of 768 records of patients from Phoenix, Arizona. It is under the continuous analysis by National Institute of Diabetes, Digestive and Kidney Disease. This dataset contains the medical profile of the patients to diagnose diabetes. There are 9 attributes in the database. All the patients are females about 21 years old. The followings are the attributes with description of the diabetes dataset.

- Class Variable (0 – Positive (or) 1- Negative)
- No. of Pregnancy
- Glucose (Oral Glucose – 2 Hours)
- Blood Pressure (mm Hg)
- Skin Thickness (mm)
- Insulin ( $\frac{mu\ U}{ml}$ ) – 2 Hours
- BMI ( $\frac{Weight\ (kg)}{Height\ (m)^2}$ )
- Function of Diabetes Pedigree
- Age (Yrs.)

### A. Kaiser-Meyer-Olkin - Sampling Adequacy Measure:

It measures the adequacy of sampling. It should be more than half (0.5) and to proceed [6].

### A. Factor Analysis of the Dataset

The Prima Indian Diabetes is analyzed with the use of factor analysis concepts.

*1) Correlation Analysis of the Dataset*

The following table 2 is presented the correlation among variables in the form of matrix.

| | | Class Variable | No. Of Pregnancy | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Function of Diabetes Pedigree | Age |
|---|---|---|---|---|---|---|---|---|---|---|
| Correlation | Class Variable | 1.000 | .222 | .467 | .065 | .075 | .131 | .293 | .174 | .238 |
| | No. Of Pregnancy | .222 | 1.000 | .129 | .141 | -.082 | -.074 | .018 | -.034 | .544 |
| | Glucose | .467 | .129 | 1.000 | .153 | .057 | .331 | .221 | .137 | .264 |
| | Blood Pressure | .065 | .141 | .153 | 1.000 | .207 | .089 | .282 | .041 | .240 |
| | Skin Thickness | .075 | -.082 | .057 | .207 | 1.000 | .437 | .393 | .184 | -.114 |
| | Insulin | .131 | -.074 | .331 | .089 | .437 | 1.000 | .198 | .185 | -.042 |
| | BMI | .293 | .018 | .221 | .282 | .393 | .198 | 1.000 | .141 | .036 |
| | Function of Diabetes Pedigree | .174 | -.034 | .137 | .041 | .184 | .185 | .141 | 1.000 | .034 |
| | Age | .238 | .544 | .264 | .240 | -.114 | -.042 | .036 | .034 | 1.000 |

Table 2. Correlation Matrix for Diabetics Patients

The above Table 2 shows the Pearson correlation coefficient among all pairs of variables. This correlation matrix is to check the pattern relationships. It is in the array of numbers which is in the form of rectangular and also given the correlation coefficients between single and each other variables. The values between single and itself are always one. In the correlation, the principal diagonal is 1.000 so that each variable has the perfect positive linear relationship with itself. Above and below values of the principal diagonal are same in the correlation matrix.

*2) Test of KMO and Bartlett:*

The following table 3 is described test of KMO and Bartlett Test of the dataset.

| | | |
|---|---|---|
| KMO - (Measures of Sampling Adequacy) | | .618 |
| Bartlett's Test - Sphericity | (Approximate Chi-Square) | 1223.570 |
| | DF | 36 |
| | Sig. | .000 |

Table 3. KMO Test and Bartlett Test for Diabetics

Normally 0 < KMO Value < 1

If KMO value is more than half (0.5), the sample is also an adequate. The KMO value is 0.618 that indicates adequate so we can proceed the Factor Analysis.

### B. Bartlett's Test of Sphericity:

It indicates the strength of relationship between the variables.

95% of the significance level $\alpha = 0.05$

If p-value /Sig is $0.000 < 0.05$, then the Analysis is valid.

The correlations between the variables are all zero.

If $p < \alpha$ then

To reject the null Hypothesis ($H_0$.)

To accept the alternate Hypothesis ($H_1$.)

So statistically, there may be an interrelationship between variables in significant manner.

The approximate value of chi-square is 1223.570 with 36 degrees.

For further analysis, Factor Analysis is an appropriate technique.

### C. Communalities:

The following table 4 is presented the communalities of the dataset.

| Variables | Initial | Extraction |
|---|---|---|
| Class Variable | 1.000 | .623 |
| No. Of Pregnancy | 1.000 | .634 |
| Glucose | 1.000 | .637 |
| Blood Pressure | 1.000 | .660 |
| Skin Thickness | 1.000 | .676 |
| Insulin | 1.000 | .513 |
| BMI | 1.000 | .521 |
| Function of Diabetes Pedigree | 1.000 | .279 |
| Age | 1.000 | .705 |

Tabel 4 Communalities for Diabetics

If communality of the value should be more than 0.5, then proceed the further step for factor analysis otherwise these variables are removed from the further step of factor analysis. The above table is showing that all the variables are above 0.5 except Function of Diabetes Pedigree variable. So, we can proceeds the further step for factor analysis [7].

### D. Total Variance Explained:

The below table 5 is discussed about the explanation of the total variance of the dataset.

| Components | Initial Eigen Values | | | Extraction of Sum of Squared Loadings | | | Rotation of Sum of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2.353 | 26.139 | 26.139 | 2.353 | 26.139 | 26.139 | 1.814 | 20.156 | 20.156 |
| 2 | 1.774 | 19.715 | 45.853 | 1.774 | 19.715 | 45.853 | 1.720 | 19.107 | 39.263 |
| 3 | 1.120 | 12.447 | 58.300 | 1.120 | 12.447 | 58.300 | 1.713 | 19.037 | 58.300 |
| 4 | .882 | 9.799 | 68.100 | | | | | | |
| 5 | .845 | 9.385 | 77.485 | | | | | | |
| 6 | .735 | 8.165 | 85.650 | | | | | | |
| 7 | .488 | 5.427 | 91.077 | | | | | | |
| 8 | .418 | 4.646 | 95.723 | | | | | | |
| 9 | .385 | 4.277 | 100.000 | | | | | | |

Table 5 Explanation of Total Variance

The Eigen value reflects the number of extracted factors from number of items that are included in the factor analysis [8]. The above table is divided into 3 segments like Sum of Squared Loadings Extraction, Initial Eigen values, and Sum of Squared Loadings. The first three variables' values are 26.139 % of variance, 19.715 % of variance and 12.447 % of variance. So the first three components are taken for further analysis. The remaining variances are not significant.

### E. Component Matrix:

In this component matrix, it is presented the correlation of the component matrix about the variables in the dataset. The table 6 contains component loading and also correlations between variable and component. The -1 to +1 are the possible correlation values range.

| Variables | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Glucose | .670 | .127 | -.414 |
| Class Variable | .638 | .206 | -.418 |
| BMI | .609 | -.280 | .267 |
| Insulin | .516 | -.473 | -.150 |
| Function of Diabetes Pedigree | .364 | -.233 | -.302 |
| Age | .427 | .710 | .133 |
| No. Of Pregnancy | .331 | .703 | .174 |
| Skin Thickness | .471 | -.597 | .312 |
| Blood Pressure | .461 | | .666 |

Table 6 Component Matrix for Diabetics

It shows how each of the items correlates with each of 3 retained components in the analysis. Here, the negative and positive correlations carry the same weight. The correlations values are less than the 0.3 or less which are not meaningful.

### F. Rotated Component Matrix:

In the below table 5, it is displayed correlation of rotated component matrix about variables in the dataset.

| Variables | Components | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Age | .828 | .140 | |
| No. Of Pregnancy | .794 | | |
| Glucose | .261 | .750 | |
| Class Variable | .316 | .723 | |

| | | | |
|---|---|---|---|
| Insulin | -.253 | .506 | .438 |
| Function of Diabetes Pedigree | -.144 | .489 | .136 |
| Skin Thickness | -.251 | .157 | .767 |
| Blood Pressure | .420 | -.161 | .676 |
| BMI | | .257 | .671 |

Table 7 Rotated Component Matrix of Diabetics

The rotated component matrix shows that Insulin, Function of Diabetes Pedigree and Skin Thickness have negative loadings in the first component and positive loading in the second and third components. The Blood Pressure is positive loading in the first and third components but negative loading in the second component [9].

### G. Component Transformation Matrix:

The component transformation matrix is shown the component correlation matrix prior and after rotation.

| Components | 1 | 2 | 3 |
|---|---|---|---|
| 1 | .394 | .695 | .602 |
| 2 | .876 | -.087 | -.474 |
| 3 | .277 | -.714 | .643 |

**Table 8 Diabetics Component Transformation Matrix**

## IV. RESULT AND DISCUSSION:

The components and eigenvalues are in X-axis and in Y-axis respectively in the Scree Plot. The graph of figure 1 is to determine and to retain the factors. It is used to find the points in the curve where it is to start flatten. The curve begins to flatten between 3 and 4. The eigenvalue is less than 1.0 component from 4 to end component. So, three factors the above from 4 are removed from the factors [10].

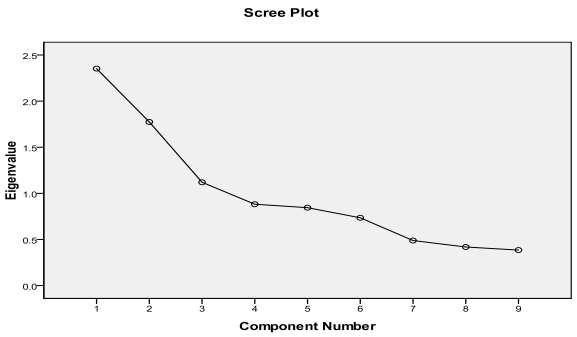


**Fig. 1. Scree Plot**

## V. CONCLUSION:

This paper includes the Dataset of Pima Indians Diabetes from the UCI repository of the machine learning databases for prediction of diabetics using factor analysis. The diabetes is increasing now a day's among the people of young adults and old age. All parameters are analyzed for the diabetes risk of a person. So prediction of diabetics is so important for diagnosis and also for providing treatment. Based on the analysis, the patient can maintain or control the level disease severity and avoid the severe effect on the patient's organ like Heart, Kidney, and Eye etc. This factor analysis is giving a solution about parameters like how it is related with other parameter of the dataset in each step model.

## REFERENCES


1. Jyoti Soni, Ujma Ansari and Dipesh Sharma, "Predictive Data Mining for Medical Diagnosis : An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975-8887), Vol. 17 – No.8, Pg.No: 43 - 48,March 2011.
2. Gunasekar Thangarasu and Dominic.D.D, "Prediction of Hidden Knowledge from Clinical Database using Data Mining Techniques", 978-1-4799-0059-6/13, IEEE.
3. Dhomse Kanchan B and Mahale Kishor M, "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", International Conference on Global Trends in Signal Processing, Information Computing and Communication, 978-1-5090-0467-6/16, Pg.No: 5-10, IEEE .
4. Gauri D.Kalyankar, Shivananda R.Poojara and Nagaraj V. Dharwadkar, "Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), Pg.No: 619-624, 978-1-5090-3243-3/17 IEEE.
5. Philip Hyland, "An Introduction to Factor Anaysis", www.philiphyland.webs.com
6. http://www.statisticshowto.com/kaiser-meyer-olkin/
7. Priya Chetty and Shruti Datt , "Interpretation of Factor Analysis using SPSS", Feb 5, 2015. https://www.projectguru.in/publications/interpretation-of-factor-analysis-using-spss/
8. Factor Analysis http://www.sjsu.edu/people/james.lee/courses/JS203/s1/Online_4.pdf
9. Exploratory Factor Analysis and Principal Components Analysis https://tandfbis.s3.amazonaws.com/rt-media/pdf/9781848729995/IBM_SPSS_5e_Chapter_4.pdf
10. Robin Beaumont, "An Introduction to Principal Component Analysis and Factor Analysis", 23, April 2012.


## AUTHORS PROFILE

**Dr. S. Kavitha** is an Assistant Professor in SRM Institute of Science and Technology, Chennai, Tamil Nadu, India. Her area of interest in research includes Data Mining, Parallel and Distributed Mining, Load Balancing, High Performance Computing and Artificial Intelligence. She presented research papers in conferences and also published papers in the International Journals.

**Mrs. E. Srividhya** received her B.Tech from Prince Shri Venkateshwara Padmavathy Engg College. Got her master degree from Vinayaka Missions Research Foundation. She joined as an Assistant Professor in the Department of Information Technology in 2010 doing her PhD at Bharath Institute of Higher Education & Research, in "Diagnosis of Diabetes by Tongue Analysis using Image Processing". Here trust area of research includes Image Processing, Deep Learning and Data Mining. Besides she is a life member in Indian Society for Technical Education and International Association of Engineers. She is engaged in teaching and research work. Despite teaching the students, she has filed a patent "Smart Diagnosis System for Diabetes" in 2018, and has 10 publications in International and national journals. Currently she is working as an Assistant Professor in Department of Computer Science and Engineering, Aarupadai Veedu Institute of Technology, Chennai.

**Mr. S. Muthuselvan,** M.E., (Ph.D) currently working as Assistant Professor Gr. II, Aarupadai Veedu Institute of Technology an ambit institution of Vinayaka Mission's Research Foundation (Deemed to be University), Tamil Nadu, India. Published more than 13 national and international journal and organizing committee for four international conference, two national conference and 11 years of teaching experience with 6 years of research experience. He is a member in following professional societies: CSI and MISTE.