Diabetes Kaggle Dataset Adequacy Scrutiny using Factor Exploration and Correlation

Viswanatha Reddy Allugunti, Elango N M, Kishor Kumar Reddy C

Abstract: Diabetes is one in everything about principal serious incessant sicknesses. In accordance with American sickness Association (ADA), it forces a gigantic financial weight at the nations. Consideration charges, in view of polygenic infection, represent 11% ($465 billion) of the complete consideration costs in the worldwide in 2011. By 2030, this assortment is anticipated to surpass $595 billion. Around the world, there are around 366 million people with polygenic disorder and it's measurable that 552 million will be influenced by 2030 comes that diabetes is that the seventh driving intention in mortality in 2030. The way to diminish these dangers could be a significant trouble. This paper destinations to demonstrate that the chose dataset with 15000 data is alright for experimentation misuse Kaiser-Meyer-Olkin. It conjointly objectives to demonstrate that grid is accomplice unit network abuse Bartlett's trial of Sphericity.

Index Terms: Diabetes. Correlation Analysis, Factor Analysis, Health care, ***Kaiser-Meyer-Olkin Measure***

1. INTRODUCTION

Diabetes is an endless issue and a main open wellness task universal. It happens while a body can't respond or outgrowth well to insulin, which is expected to keep up the charge of glucose. Diabetes might be controlled with the assistance of insulin infusions, a refreshing nourishment routine and customary exercise however there is no entire treatment is accessible. Diabetes results in significantly other sickness, for example, visual impairment, blood pressure, heart issue, and kidney issue and nerve hurt. There are 3 high types of diabetes mellitus: Type 1 Diabetes Mellitus outcomes from the body's inability to give insulin. This structure become earlier known as insulin-based diabetes mellitus.Type2 Diabetes Mellitus end from insulin opposition that is a condition wherein cells neglect to utilize insulin pleasantly, notwithstanding the way that for now and again likewise with a flat out insulin inadequacy. This sort was earlier called non insulin-based diabetes mellitus. Gestational diabetes is the 1/3 preeminent shape and happens while a pregnant women in the past shows up anticipation of diabetes extend a high blood glucose level. So as to robotize the general technique for diabetes expectation and seriousness estimation, diabetic database is needed.

This paper destination to demonstrate that dataset is adequate for experimentation by utilizing Kaiser-Meyer-Olkin Measure. Further, viewpoint investigation is executed to a lot of watched factors that tries to find basic components from which the discovered factors had been created. Further. Bartlett's trial of sphericity is done to check your connection framework is a distinguishing proof grid, which may show that your factors are random and thusly unacceptable for structure recognition. Further, relationship assessment is done that estimates the vitality of relationship between two factors and the course of the association.

The unwinding of the paper is set up as pursues: Chapter 2 delineates the writing canvases identified with diabetes expectation, section three gives dataset description, factor analysis and correlation analysis and part 4 finishes up the paper seen by utilizing references.

2. RELEVANT WORK

Gyorgy J. Simon, Pedro J.Caraballo, et al., [1] proposed the methodology of distributional affiliation principle mining to choose units of danger components and the relating quiet subpopulations which are altogether stretched out risk of advancing to diabetes. What's more, to discover sets of risk issue, directly here utilizations posterior up outline calculation which produces most extreme reasonable rundown that portrays subpopulations at high danger of diabetes. The Subpopulation analyzed through this synopsis included most high danger of sufferers, had low cover and have been at exceptionally high threat. This system is utilized for while the patient having unreasonable threat.

Dr. Zuber khan, shaifali sing, et al., [2] worked at the idea of Diabetes Mellitus the utilization of k-Nearest Neighbor set of guidelines that is greatest Important system of Artificial Intelligence. The precision expense is showing that what number of yields of the information of the test dataset are same as the yield of the realities of various capacities of the talented dataset. The bumbles charge is locating that what number of yields of the records of the test dataset aren't same as the yield of the measurements of different abilities of the training dataset. The outcome they affirmed that as the expense of alright builds, precision charge and blunder rate will development. K-Nearest Neighbor set of standards is one of the most extreme basic methods of AI that is utilized broadly for demonstrative capacities. Through KNN additional Accurate outcomes might be get. This methodology could be successful for the tutoring actualities set which is huge.

Dr.PramanandPerumal and Sankaranarayanan [6] proposed a thought regarding diabetes mellitus its finding

Revised Manuscript Received on June 10, 2019.

Viswanatha Reddy Allugunti, Research Scholar, VIT University, Vellore, T.N, India
Dr. Elango NM, Professor, VIT University, Vellore, T.N, India
Kishor Kumar Reddy C, Associate Professor, Stanley College of Engineering & Technology for Women, Hyderabad, Telanagana, India.

the use of actualities mining with insignificant number of credits completed to classification calculations. They toiled on Apriori and FP-development systems. In FP-development the capricious data structure normal example tree is being done for putting away packed basic certainties roughly regular example. It is found that both of the systems create the equivalent wide assortment of continuous units as a significance equivalent number of guidelines for the indistinguishable recognized dataset under similar imperatives. With the assistance of data Apriori and FP-development calculations, the calculation cost diminishes and moreover the sort execution increments.

SatyanarayanaGandi and AmarendraKothalanka [7] worked at the underlying preparing data set to the best quality level method to remove the most solid data set, on that ideal dataset they actualized grouping with Bayesian classifier. Bayesian classifier methods is utilizes getting preparing data set and convert it into marked insights. At first they separate the most dependable list of capabilities from current tutoring realities and figures the high caliber and poor likelihood, until the new informational index whenever shaped with indistinguishable size and advances the present day produced dataset for characterization their it groups the testing dataset with new trademark.

Sanchitapaul and DilipkumarChoubey [9] proposed a methodology for trademark decision, class and utilized Genetic Algorithm, Multilayer Perceptron Neural people group on diabetes data set. With abilities decision technique the use of Genetic arrangement of principles they upgrade the exactness anyway finished somewhat less ROC. With trademark Selection procedure hereditary arrangement of standards propelled precision yet achieved considerably less ROC by method for utilizing GA,MLP NN technique class ROC is likewise ventured forward.

RamkrishnanShrikant and RakeshAgrawal proposed a deliberate structure of structure a risk expectation adaptation for sort 2 diabetes disorder. The GBRE set of guidelines recognizes the top of the line set of markers that can expect danger dimension of diabetes and afterward numerous classifiers are talented and their exactness are estimated. Alan J. Garber,MD and Martin J.Abrahamson et al., [10] propelled case watch comprises of Evalution for Complications and arranging, Lifestyle Modifications, Algorithm for including/Intensifying insulin, CVD Risk issue set of guidelines, Profiles of hostile to diabetic Medications. Standards of the AACE Algorithm for the cure of sort 2 diabetes.

Rohit Prasad Bakshi and SonaliAgrawal [16] proposed a logical system of structure a danger of forecast rendition for kind-2 diabetes illness. The GBRE set of standards finds the phenomenal arrangement of that may decided possibility level of diabetes after which a few classifiers are prepared and their exactness are being estimated. The classifier has been chosen by method for vote throwing strategy technique. The proposed methodology might be executed obviously in forecast demonstrating of various sicknesses.

S.Sapna and Dr.A.Tamilarasi [17] proposed a thought of Genetic Algorithm and Fuzzy gadget on chromosomes. To Obtained the precision of chromosome and to survey the diabetes in diabetic influenced individual GA is completed. The association between fluffy machine and hereditary arrangement of principles is bidirectional. Hereditary Algorithms are connected to manage different improvement issues involves fluffy machine. Utilizing GA enhancement of chromosome is acquired and essentially dependent on the charge of vintage masses diabetes might be obliged in new populace to get chromosomal precision.

SrideivanaiNagarajan and R.M. Chandrasekaran [18] proposed a strategy for development of conclusion of gestational diabetes with realities mining techniques. Likewise they Analyze the general execution of ID3, Naïve Bayes, C4.5, and Random tree i.E. The arrangement of standards for directed Learning. They utilized the records set of Pregnant Womens. The results they found that Random tree served to be the top notch one with better exactness and least botches cost.

VeenaV.Vijayan and AswathyRavikumar [19] talked about the rule data mining set of standards, K-Means Algorithm, Amalgam KNN set of principles and ANFIS set of guidelines They proposed the examine of Expectation Maximization calculation utilized for inspecting to decide and boost the anticipation in progressive emphasis cycles. K-Nearest Neighbor Algorithm is utilized for grouping of things and utilized for forecast of names dependent on a couple of nearest training precedents in the element region. K-Means calculation pursues segment strategies basically dependent on some enter parameters on the dataset of n devices. They referenced about Amalgam Algorithm joins each the capacity of K-Nearest Neighbor and K-Means with a couple of extra handling and the Adaptive Neuro Fuzzy Inference System which consolidates the Features of Adaptive Neural Network and Fuzzy Inference System. They choose the dataset from PIMA Indian Diabetic Set from University of California.

K.Rajesh and V.Sangeetha [20] recommended that measurements digging dating for green classification they connected certainties mining methodologies to group diabetes medicinal records and anticipate the influenced individual being influenced with diabetes or never again. They offered a gadget which gave instruction information on that insights work significance investigation is accomplished then correlation of sort set of guidelines, Selecting classifier at that point ventured forward classification calculation is executed after which decided out the appraisal that as contrasted and training records. They did C4.Five Algorithm gave type charge of ninety one%.

Dr. B .L. Shivkumar and S. Alby [21] affords a review paper for records mining strategies which have been normally completed to diabetes information assessment and forecast of confusion. They accomplished an investigation of various shows and research accomplished through various looks into. From the assessment of different examinations papers it's miles clear that the predominance of diabetes is having powerful connection with sicknesses like Wheeze Edema, Oral ailments, Female Pregnant and increment of age. Utilizing information mining techniques the risk of diabetes can be normal that is useful for early location of the malady.

Carlos Fernandez_Llatas and Antonio Martinez_Millanu [22] proposed utilizing Interactive Pattern Recognition procedures for the iterative design of conventions and breaking down the issues of the use of framework mining to conclude care streams and the best approach to adapt the resulting spaghetti Effect. Underneath parent proposes the concise depiction about techniques which recently utilized for forecast investigation.

3. RESULTS AND DISCUSSION

3.1 Dataset Description

The dataset (https://www.Kaggle.Com/fmendes/diabetes-from-dat263x-lab01) is at first from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to indicatively anticipate whether or no longer a patient has diabetes, in light of certain symptomatic estimations included inside the dataset. A few imperatives were situated on the choice of those occasions from a greater database. The datasets envelop various restorative indicator (unprejudiced) factors and one target (subordinate) variable, Outcome. Free factors comprise of the wide assortment of pregnancies the influenced individual has had, their BMI, insulin stage, age, etc. In Table 1, A recommends Pregnancies, B demonstrates Plasma Glucose, C demonstrates Diastolic Blood Pressure, D speaks to Triceps Thickness, E speaks to Serum Insulin, F proposes BMI, G demonstrates Diabetes Pedigree, H shows Age and I show Diabetic. It has 15000 data. In Table 1, J Column, 1 demonstrates Diabetes and 0 shows no diabetes.

Table 1: Dataset Description

A	B	C	D	E	F	G	H	I	J
1354778	0	171	80	34	23	43.50973	1.213191	21	0
1147438	8	92	93	47	36	21.24058	0.158365	23	0
1640031	7	115	47	52	35	41.51152	0.079019	23	0
1883350	9	103	78	25	304	29.58219	1.28287	43	1
1424119	1	85	59	27	35	42.60454	0.549542	22	0
1619297	0	82	92	9	253	19.72416	0.103424	26	0
1660149	0	133	47	19	227	21.94136	0.17416	21	0
1458769	0	67	87	43	36	18.27772	0.236165	26	0
1201647	8	80	95	33	24	26.62493	0.443947	53	1
1403912	1	72	31	40	42	36.88958	0.103944	26	0
1943830	1	88	86	11	58	43.22504	0.230285	22	0
1824483	3	94	96	31	36	21.29448	0.25902	23	0
1848869	5	114	101	43	70	36.49532	0.07919	38	1
1669231	7	110	82	16	44	36.08929	0.281276	25	0
1683688	0	148	58	11	179	39.19208	0.160829	45	0
1738587	3	109	77	46	61	19.84731	0.204345	21	1
1884264	3	106	64	25	51	29.04457	0.589188	42	1
1485251	1	156	53	15	226	29.78619	0.203824	41	1
1536832	8	117	39	32	164	21.231	0.089363	25	0
1438701	3	102	100	25	289	42.18572	0.175593	43	1

3.2 Factor Analysis

Factor Analysis is an exploratory methodology done to a firm of decided factors that looks to find fundamental elements from which the discovered factors had been created. Factor assessment is finished at the relationship framework of the decided factors. Table 2, shows the infer and across the board deviation calculations.

Table 2: Descriptive Statistics on Dataset

Descriptive Statistics			
	Mean	Std. Deviation	Analysis N
Pregnancies	3.22	3.391	15000
Plasma Glucose	107.86	31.982	15000
Diastolic Blood Pressure	71.22	16.759	15000

Triceps Thickness	28.81	14.556	15000
Serum Insulin	137.85	133.068	15000
BMI	31.5096460410	9.75899973405	15000
Diabetes Pedigree	.39896774896	.377943532154	15000
Age	30.14	12.090	15000
Diabetic	.33	.471	15000

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy is a measurement that demonstrates the extent of change to your factors that is likely coming about because of fundamental elements. High qualities (close 1.0) by and large propose that an issue assessment might be valuable with your measurements. On the off chance that the charge is substantially less than 0.50, the results of the thing examination potentially probably won't be valuable. For the given data the KMO rating is 0.609, appeared Table 3. This KMO worth demonstrates that the example turned out to be alright and therefore attractive, and the circulation of expense is alright for directing perspective assessment.

Bartlett's trial of sphericity tests the hypothesis that your connection network is a personality lattice, which could propose that your factors are inconsequential and thus inappropriate for structure location. Little qualities (under 0.05) of the criticalness degree demonstrate that a viewpoint assessment can be valuable with your data. The Bartlett's Test of Sphericity charge changed into exceedingly extensive (Chi square = 7416.714, $p < 0.001$), and along these lines thing examination is proper, demonstrated in Table 3.

Table 3: KMO and Bartlett's Test on Dataset

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.609
Bartlett's Test of Sphericity	Approx. Chi-Square	7416.714
	df	36
	Sig.	.000

Table 4 clarifies the Eigen esteems identified with each straight part (thing) before extraction, after extraction and after turn. Before extraction it has analyzed 9 direct parts. The Eigen esteems identified with each issue establish the difference clarified through that specific direct component. It additionally introductions the Eigen esteems in expressions of percent of fluctuation. We ought remember all components with Eigen esteems mutiple. The principal thing inside the Table 3 clarifies 20.597 % of the absolute fluctuation. Additionally, the second one issue parts clarify 32.139 % of the entire difference. Through thing investigation, two premier parts have been separated from the 9 factors.

Table 4: Total Variance using Principal Component Analysis

Total Variance Explained									
	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.882	20.912	20.912	1.882	20.912	20.912	1.854	20.597	20.597
2	1.010	11.228	32.139	1.010	11.228	32.139	1.039	11.542	32.139
3	.992	11.024	43.164						
4	.977	10.851	54.014						
5	.963	10.703	64.717						
6	.940	10.450	75.167						
7	.907	10.075	85.242						
8	.860	9.550	94.792						
9	.469	5.208	100.000						
Extraction Method: Principal Component Analysis									

Table 5 demonstrates the turned part grid utilizing the extraction technique for Principal Component Analysis. Every one of the components having different qualities is assembled under the age wherein it has the most astounding cost. Table 5 and Fig. 1 gives the 2 issue added substances as got from the varimax turn technique for segment assessment, each given an 'interpretative' call.

Table 5: Rotated Component matrix

Component Matrix[a]		
	Component	
	1	2
Pregnancies	.612	.033
Plasma Glucose	.219	.460
Diastolic Blood Pressure	.169	-.172
Triceps Thickness	.259	.612
Serum Insulin	.417	-.121
BMI	.347	-.080
Diabetes Pedigree	.272	-.611
Age	.540	.002
Diabetic	.839	-.009
Extraction Method: Principal Component Analysis.		
a. 2 components extracted.		

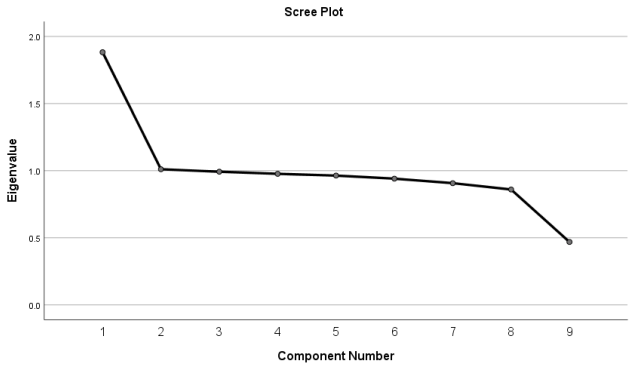

Fig. 1: Screen Plot of Components on Rotation Matrix

3.3 Correlation Analysis

Correlation is a bivariate examination that estimates the vitality of relationship between two factors and the course of the association. In expressions of the vitality of pursuing, the estimation of the connection coefficient fluctuates among +1 and - 1, demonstrated in Table 6. A cost of ± 1 demonstrates an incredible confirmation of relationship among the 2 factors. As the connection coefficient expense is going toward zero, the relationship among the 2 factors can be flimsier. The bearing of the association is shown by methods for the sign of the coefficient; a + sign proposes a wonderful dating and a – signal recommends a horrendous relationship. In Table 6, PR showed Pearson Correlation, S demonstrates sig. 2 followed, A demonstrates Pregnancies, B proposes Plasma Glucose, C recommends Diastolic Blood Pressure, D speaks to Triceps Thickness, E speaks to Serum Insulin, F indicates BMI, G demonstrates Diabetes Pedigree, H recommends Age and I demonstrates Diabetic.

Table 5: Correlations Analysis on Dataset

Correlations										
		A	B	C	D	E	F	G	H	I
Pregnancies	PR	1	.055**	.044**	.064**	.104**	.086**	.054**	.137**	.407**
	S		.000	.000	.000	.000	.000	.000	.000	.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
Plasma Glucose	PR	.055**	1	.007	.027**	.034**	.021*	.009	.039**	.128**
	S	.000		.377	.001	.000	.011	.267	.000	.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
Diastolic Blood Pressure	PR	.044**	.007	1	.011	.023**	.016	.014	.041**	.091**
	S	.000	.377		.174	.006	.052	.084	.000	.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
Triceps Thickness	PR	.064**	.027**	.011	1	.030**	.025**	-.001	.061**	.153**
	S	.000	.001	.174		.000	.002	.907	.000	.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
Serum Insulin	PR	.104**	.034**	.023**	.030**	1	.051**	.046**	.088**	.247**
	S	.000	.000	.006	.000		.000	.000	.000	.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
BMI	PR	.086**	.021*	.016	.025**	.051**	1	.029**	.063**	.211**
	S	.000	.011	.052	.002	.000		.000	.000	.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
Diabetes Pedigree	PR	.054**	.009	.014	-.001	.046**	.029**	1	.056**	.170**
	S	.000	.267	.084	.907	.000	.000		.000	.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
Age	PR	.137**	.039**	.041**	.061**	.088**	.063**	.056**	1	.343**
	S	.000	.000	.000	.000	.000	.000	.000		.000
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
Diabetic	PR	.407**	.128**	.091**	.153**	.247**	.211**	.170**	.343**	1
	S	.000	.000	.000	.000	.000	.000	.000	.000	
	N	15000	15000	15000	15000	15000	15000	15000	15000	15000
**. Correlation is significant at the 0.01 level (2-tailed).										
*. Correlation is significant at the 0.05 level (2-tailed).										

4. CONCLUSIONS

This paper interests to demonstrate that dataset is adequate for experimentation by utilizing Kaiser-Meyer-Olkin Measure. For the given information the KMO rating is 0.609, proposes that the example transformed into satisfactory and therefore fit, and the appropriation of expense is adequate for achieving component investigation. Further, viewpoint assessment is connected to a fixed of decided factors that tries to find hidden components from which the found factors have been produced. Further. Bartlett's trial of sphericity is cultivated to check your connection grid is a distinguishing proof network, which may show that your factors are inconsequential and

therefore unacceptable for shape discovery. The Bartlett's Test of Sphericity cost turned out to be colossally full-estimate (Chi square = 7416.714, $p < 0.001$), and thusly factor assessment is proper. Further, relationship assessment is done that estimates the power of association between two factors and the way of the association.

REFERENCES

1. GyorgyJ.Simon,PedroJ.Caraballo,Terry M. Therneau,Steven S. Cha, M. Regina Castro and Peter W.Li "Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus," *IEEE Transanctions on Knowledge and Data Engineering* ,vol 27,No.1,January 2015
2. Dr.Zuber khan, shaifalisingh and KratiSexena," Diagnosis of Diabetes Mellitus using K- Nearest Neighbor Algorithm," *in proceeding of International Journal of Computer Science Trends and Technology*, vol.2 , July-Aug 2014
3. Mukeshkumari and Dr. RajanVohra, "Prediction of Diabetes Using Bayesian Network,"*in proceeding of International Journal of Computer Science and Information Technologies*, vol. 5 , 2014
4. Jianchao Han, Juan C. Rodriguze and Mohsen Beheshti , "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner,"*in proceeding of Second International Conference on Future Generation Communication and Networking*, vol.2,2008
5. Wang ZuoCheng and XUE Li Xia ,"A Fast Algorithm for Mining Association Rules in Image," *in proceeding of International Conference on Data Engineering*, vol.5, 2008
6. Dr. PramanandPerumal and Sankaranarayanan, "Diabetic prognosis through Data Mining Methods and Techniques," *in proceeding of International Conference on Intelligent Computing Applications*, vol. 2, 2014
7. SatyanarayanaGandi and AmarendraKothalanka,"An Efficient Expert System For Diabetes By Naïve Bayesian Classifier," *in proceeding of International Journal of Engineering Trends and Technology* ,vol. 4 ,Issue 10 , Oct 2013
8. RamkrishnanShrikant and RakeshAgrawal,"Fast Algorithms for mining association rule," *in proceeding of IEEE International Conference on Data Engineering*,vol.16,2007
9. Dilip Kumar Choubey and SanchitaPaul,"GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis *",in proceedings of I.J.Intelligent System and Applications*, vol.1,pp.49-59,2016
10. Alan J. Garber,MD and Martin J.Abrahamson, Case study on "AACE/ACE Comprehensive Diabetes Management Algorithm"
11. H. S. Kim, A. M. Shin, M. K. Kim, and N. Kim, "Comorbidity study on type 2 diabetes mellitus using data mining," *in proceedings of Korean J. Intern. Med.*, vol. 27, no. 2, pp. 197– 202, Jun. 2012
12. KawitaRawat and KawitaBhurse" A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network,"*in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, vol.4, Nov. 2014
13. G. S Collins, S. Mallett, O. Omar, and L.-M. Yu, "Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting,"in proceedings of BMC Med., 9:103,Sept. 2011
14. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proceedings of 20th VLDB, Santiago, Chile, 1994
15. M. A. Hasan, "Summarization in pattern mining," in proceedings of Encyclopedia of Data Warehousing and Mining, 2nd ed. Hershey, PA, USA:Information Science Reference, 2008
16. R.P.Bakshi and S.Agrawal,"Modeling Risk of Prediction of Diabetes - a preventive Measure," in proceedings of BMC Med.,2012.
17. S.Sapna and Dr.A.Tamilarasi,"Implementation of Genetic algorithm in Predicting Diabetes" in Proceedings of International journal of Computer science ,vol.9,Issue.1,N0.3,Jan-2012.
18. S. Nagarajan and R.M.Chandrasekaran,"Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes" in proceedings of International Journal of Current Research and academic Review, vol. 2,No. 10,pp. 91-98.
19. V.Vijayan and A.Ravikumar," Study of data mining algorithms for Prediction and diagnosis of diabetes mellitus," in proceedings of International Journal of Computer Application, vol. 9,No. 17, June 2014.
20. J.Tuomilehto, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impared glucose tolerance",in proceedings of International Journal of Medical Research, vol. 344,no. 18,pp. 1343-1350, 2001.
21. K.Rajesh and V.Sangeetha,"Application of Data Mining Methods and Techniques for Diabetes Diagnosis," in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, September 2012.
22. B.L. Shivkumar and S Alby,"A Survey on Data Mining Technologies for Prediction and Diagnosis of Diabetes," in proceedings of International Conference on Intelligent Computing Application, 2014.
23. Carlos Fernandez_Llatas and Antonio Martinez_Millanu, "Diabetes care related process modelling using Process Mining Techniques Lessons Learned in the Application of Interactive Pattern Recognition :Coping with the Spaghetii Effect, 2015