Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.8, No.4, 2017

Deep Insight into Diabetic Data with the Help of Association Rule Mining

Adil Aziz [1] Miss Asma Sajid [1] Shahbaz Ahmad[2]
1.Department of Computer Science, Govt. College University, Faisalabad
2.Department of Computer Science, National Textile University, Faisalabad

Abstract
Diabetes is most emerging chronic disease in Pakistan as well as around the globe. It may be in form of insulin or glycogen but both forms are hazardous for patients. A decades before diabetic patients were normally mature person whose age range was forty five to onward. But now day youngsters are also indulging in it. Diabetes is associated with mental sickness and strain while these attributes varies from culture to culture and race to race through the world. Developing countries are normally facing major issues like health and poverty along with less financial sources. Whereas most of the people are living nonstandard life and always facing trouble and tension to run their daily lives. Poverty, tension, stress and strain are home hub for diabetes. Data mining technique is very useful for retrieving the association of multiple attributes in a given massive database. In this way we can identify the major attribute among the millions of records which cause the occurrence of diabetic.
Keywords: Diabetic, Association Rule, Data Mining, Apriori

Introduction

Diabetes is currently positioned at the 6th driving reason for death by sickness in the U.S (National diabetes reality sheet, Atlanta2004). Its treatment and in addition the administration of diabetes related inconveniences remains a top need for governments around the world, since the monetary weight in 2007 alone surpassed $174 billion[1].

Diabetic patients are increasing gradually in Pakistan as well all over the globe. This is a chronic disease which doesn't have a permanent remedy. Patients can just have some precautionary measures so that he can stand his sugar level in the stated limit. If the level of sugar is higher in the patient then it can have some harmful effect on patient's health. I have taken data set of almost one lake plus patients from the UCI repository. This data carries multiple attributes concern to diabetes. Attributes comprised of different test related to sugar like serum test A1C, metformin, chlorpropamide, glyburide, repaginate, nateglinide, pioglitazone test and many more. Patients are grouped with respect to their age its range is from one year to hundred years as Caucasian nations' data set.

Normally diabetic disease emerges when an individual feel stress and strain and age is above forty-five may be some other causes that may be associated with this disease .It is troubles and boring to define association of attributes from this data set which are main sources of this disease. Association of major attributes from this data set that causes diabetic then it will be helpful for health care agencies to use this information for future planning and decision making. (Wang et al., 2012). Selection of the specified attributes that need to drive the require information is made. Then allocated the data set with respect to Caucasian nations so that each patient can be observed individually. It is possible to find result that varies from nation to nation or may have the same as derived from single nation. Most of the time it is observed that mature age group like forty plus are target of this chronic disease. But on the other hand this chronic is also scattering among youngsters. The DATASET is taken from UCI Repository of nearly 75000 patients and to mine these attributes to carry out our research. Data Set contains many attributes of each patient. Different types of categories are made on the basis of their Gender, Their Age Group, Glucose Test Report Value and A1C Test result. Association Rule Mining Algorithms are used here for analysis of this diabetes data.

Race

Race as a social build, is a gathering that have comparable physical qualities and distinctive individuals. Initially, it is utilized to allude to a typical dialect of the general population, then express alliance nation, began his vocation in the seventeenth century to the physical qualities (ie phenotypes). The term is more often than not in the feeling of a general scientific classification of utilization, from the nineteenth century; the distinctive gatherings spoke to by the phenotypes characterized. Social states of mind and profession bunch change after some time, as the premise for open cooperation taxonomy meaning of familiarity with the key sorts of people. Researchers considered old organic nature, are for the most part not supported in physiological and behavioral attributes of aggregate racial separation clarification.

Material and Method

A1C Test

A1C test is a blood test; the normal blood glucose level gives data about the individual, likewise called glucose, the data in the most recent three months. The A1C test depends on glucose hemoglobin, oxygen-conveying protein connection.

Glucose Serum

If you have a fasting blood glucose tests: horizontal 100~125 mg / dL means you have impaired fasting glucose, a type of pre-diabetes. This increases your risk of type 2 diabetes. 126 mg / dL and above A level usually means that you have diabetes.

Age

This attribute represents the age of the patients both male and female. The database comprised of patients with all ages ranging from one year to ninety years. For simplification and efficient results we develop groups of patients with respect to their age.in this research study age may help us for recognition of factors that coordinate with diabetic to attack the patient. Normally it is seen that patients with diabetic has low immunity in their genetic system. And the patient with age over fifty naturally have low immune system.so age factor can prove more significant for novel results.

Gender

This attribute reflect the information about the sex specification of patients. Just male and female are included in the database.

Step-1

The step one includes the data selection. Data is selected from the UCI repository provided list of patient examined in different hospitals. Dataset include data of different nations so sort the data according to their nation characteristic. This research study based on Caucasian nation dataset so the remaining data is excluded from the dataset. Different tables are developed according to the value of their diabetes related tests and their Gender and age. Male and Female tables are represented separately.

Step-2

Association analysis is performed to calculate confidence and support. These both factors help to extract the more associated attributes among the tables.

The Graphical Analysis Result examination was performed by using R language as a dialects tool. R is a programming normally use for data analysis in proficient way. Different types of queries are used to perform graphical operation by R language.

R tool is a statistical analysis tool used here for analysis of the dataset.

Literature Review

In operation, the quick calculation for mining affiliation rules in vast database [2], the creators propose a calculation known from the earlier, to be found in huge scale, essentially value-based, deals affiliation rules database. The calculation is found in already known arrangement of calculations and Mining Development Association Law. Wellbeing data, a great deal of work as of now utilize information mining already just business applications. The principle motivation behind which has been snared in a great deal of matches with the side effects of patients based learning frameworks conclusion. It is hard to instigate from among an arrangement of solid demonstrative tenets manifestations might be organized in an unending, on the grounds that the subsequent theory may have unacceptable forecast exactness[3].

Notwithstanding, different scientists have been utilizing affiliation standards to enhance the level estimate at 90 for every penny guaranteed to concoct further refinements [4] by its mix with regulated learning technique. The scientists connected their work on malignancy, yet they guarantee this could be stretched out to the finding of different sicknesses. Affiliation examination, it is additionally known not been utilized to likelihood explanations, for example, "if the patient is accepting treatment A, there is a likelihood of 0.35, they will indicate side effects Z" [5]. Building up connections influence the utility of a specific patient treatment arrangement, which might be valuable.

Class affiliation rules (CARs) are fundamentally used to assemble prescient order model; they can likewise be utilized to portray the connection between thing sets and classes between the imprints. The last is extremely famous in restorative information mining. For instance, disease transmission specialists frequently seen as confirmation that danger components (thing set), and the relationship between the test outcomes DIBATES guideline (class name) between. In any case, a subset of this present reality, the end client is normally

intrigued by the related class rules. Specifically, they may consider just contain things premise set from an arrangement of client characterized principles of no less than one thing guideline set. For instance, when high-hazard bunches in which HIV contamination grouping, the study of disease transmission tends to concentrate on incorporate demographic data, for example, sexual orientation, age, and conjugal status standard in the guideline premise [6].

After two innocent procedures is set by applying the term limitation to pretreatment or preparing ventures to determine this issue. In any case, this strategy is time-concentrated. Thusly, this paper presents requirement into the class affiliation standard mining process powerful way. Exploratory results demonstrate that mining time and memory utilization of the calculation is better than two fundamental techniques. The genuine advantage of our technique by genuine DIBATES area of use showing.

In 2015 associate rule (CTFI), was planned that remodel original dealing knowledge into new smaller dealing knowledge with all data of frequent item sets. By this omit the information with given transactions, got a replacement merge transactions and so the frequent thing sets are often obtained from building count Table of all items in new merge transactions, (CTFI) uses Intersection operation to come up with frequent item sets supported Count Table that compresses the things. The Bitwise-And operation is way quicker than the normal item examination methodology utilized in several Apriori-like rule. CTFI rule is that the extension of enhancements unlimited in ARM. Affiliation rules decided Present in light of a legitimate concern for the utilization of a few measures intense social database (typically taking into account least backing and least certainty).

The evaluation stage enables the comparison of models and results from any data mining model by using a common yardstick, such as lift charts, profit charts, or diagnostic classification charts. Finally, deployment relates to the actual implementation and operationalization of the data mining models. Data mining techniques can be broadly classified based on what they can do, namely description and visualization; association and clustering; and classification and estimation, which is predictive modeling [8].

REFERENCES

[1] T. Dall, S. E. Mann, Y. Zhang, J. Martin, Y. Chen, P. Hogan, and M. Petersen, “Economic costs of diabetes in the U.S. in 2007,” *Diabetes Care*, vol. 31, no. 3, pp. 596–615, 2008.

[2] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” *J. Comput. Sci. Technol.*, vol. 15, no. 6, pp. 487–499, 1994.

[3] A. Rajak and M. K. Gupta, “Association rule mining-applications in various areas,” *Proc. Int. Conf. Data Manag. Ghaziabad, India*, pp. 3–7, 2008.

[4] R. C. Basole, M. L. Braunstein, V. Kumar, H. Park, M. Kahng, D. H. orng P. Chau, A. Tamersoy, D. A. Hirsh, N. Serban, J. Bost, B. Lesnick, B. L. Schissel, and M. Thompson, “Understanding variations in pediatric asthma care processes in the emergency department using visual analytics,” *J. Am. Med. Inform. Assoc.*, vol. 22, no. 2, pp. 318–323, 2015.

[5] H. C. Koh and G. Tan, “Data mining applications in healthcare,” *J Heal. Inf Manag*, vol. 19, no. 2, pp. 64–72, 2005.

[6] D. Nguyen, B. Vo, and B. Le, “CCAR: An efficient method for mining class association rules with itemset constraints,” *Eng. Appl. Artif. Intell.*, vol. 37, no. January, pp. 115–124, Jan. 2015.

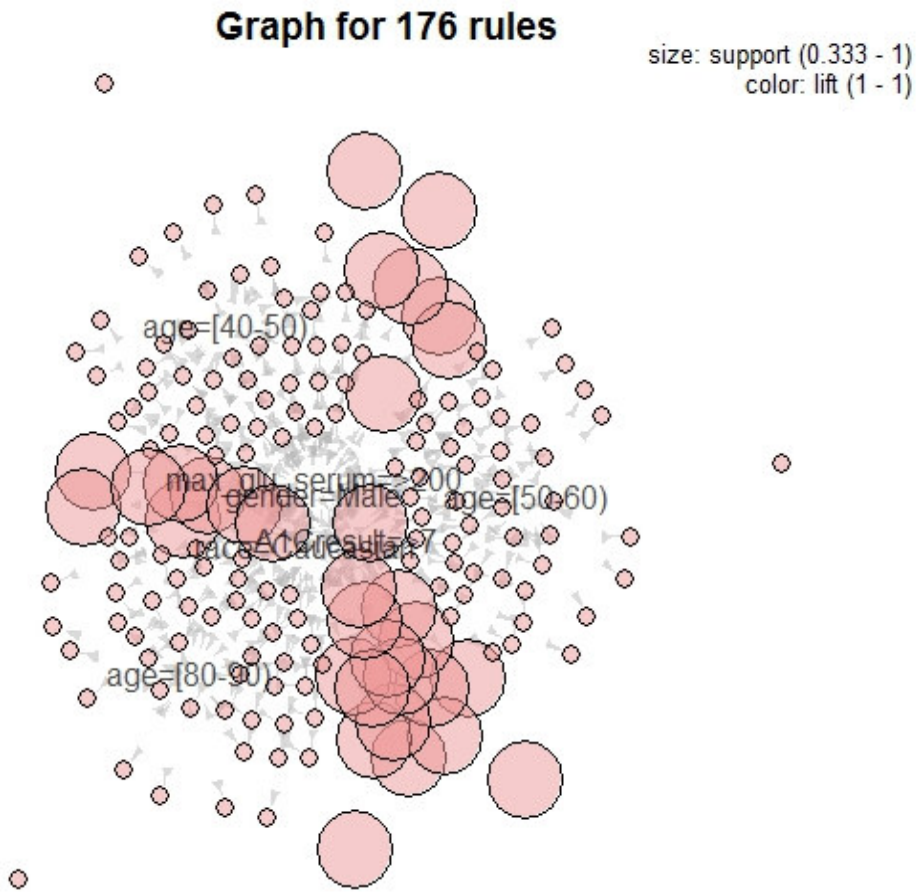
[8] M. Farahmandian, Y. Lotfi, and I. Maleki, “Data Mining Algorithms Application in Diabetes Diseases Diagnosis : A Case Study,” vol. 3, no. 1, pp. 989–997, 2015.

Let consider the given tables.

Race	A1C result	Max_glu_serum	Age	Gender
Caucasian	>7	>200	[80-90)	Male
Caucasian	>7	>200	[40-50)	Male
Caucasian	>7	>200	[50-60)	Male

Table: 3 Caucasian (7, 200) Male

This Table contains data of Caucasian race accomplished of Male gender whose both tests like A1C Result > 7 and Max_glu_serum > 200. So this table indicates that majority of male patients whose age is 40-60 and 80-90 carries diabetes. This data has 176 association rules.



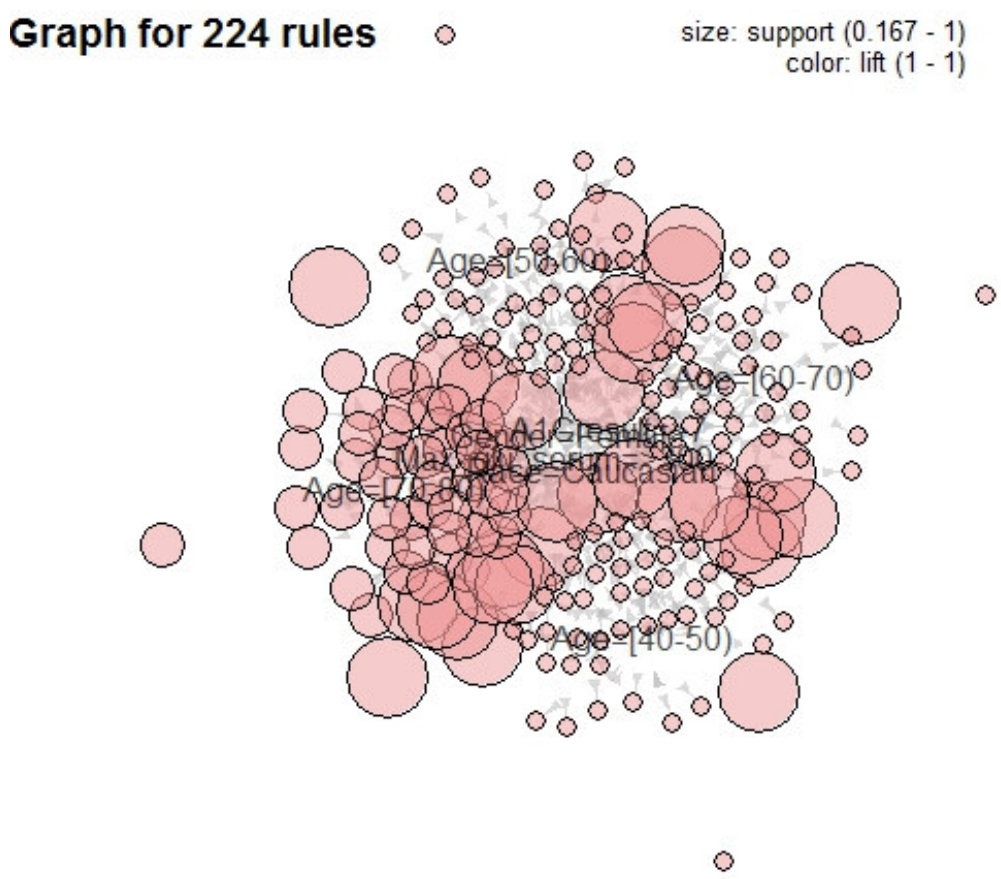
Let consider another table

Race	A1C result	Max_glu_serum	Age	Gender
Caucasian	>7	>300	[40-50)	Female
Caucasian	>7	>300	[50-60)	Female
Caucasian	>7	>300	[70-80)	Female
Caucasian	>7	>300	[60-70)	Female
Caucasian	>7	>300	[70-80)	Female
Caucasian	>7	>300	[70-80)	Female

Table: 4 Caucasian (7,300) Female

This Table contains data of Caucasian race accomplished of Female gender whose both tests like A1C Result > 7 and Max_glu_serum > 300. So this table indicates that majority of Female patients whose age is 70-80 have more chances that they carries diabetes. While from age 40-60 have some chances to carry diabetes. This data has 224 association rules.

Graph for this table is as under.



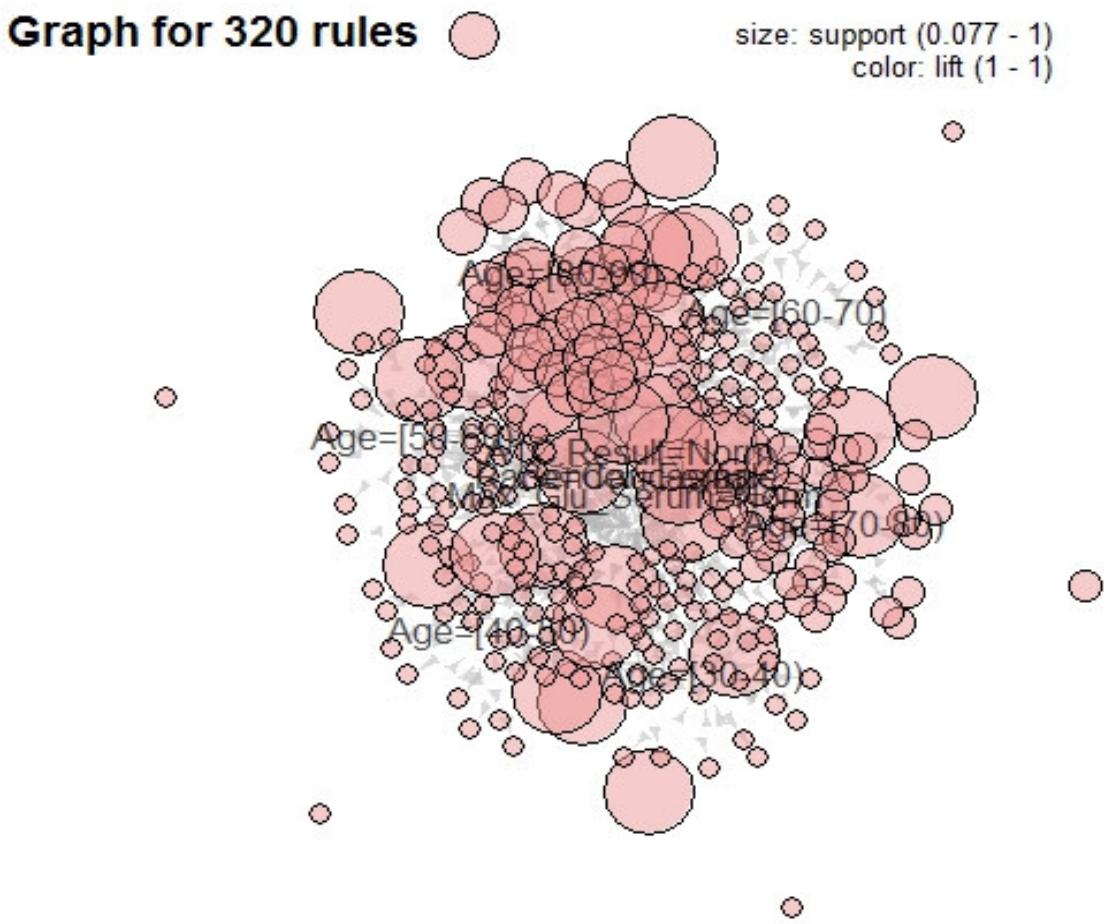
Another Table is under discussion.

Race	A1C Result	Max_Glu_Serum	Age	Gender
Caucasian	Norm	Norm	[80-90)	Female
Caucasian	Norm	Norm	[60-70)	Female
Caucasian	Norm	Norm	[80-90)	Female
Caucasian	Norm	Norm	[80-90)	Female
Caucasian	Norm	Norm	[70-80)	Female
Caucasian	Norm	Norm	[70-80)	Female
Caucasian	Norm	Norm	[50-60)	Female
Caucasian	Norm	Norm	[40-50)	Female
Caucasian	Norm	Norm	[80-90)	Female
Caucasian	Norm	Norm	[80-90)	Female
Caucasian	Norm	Norm	[80-90)	Female
Caucasian	Norm	Norm	[70-80)	Female
Caucasian	Norm	Norm	[30-40)	Female

Table: 5 Caucasian (Normal) Female

This Table contains data of Caucasian race accomplished of Female gender who's both tests like A1C Result and Max_glu_serum are in Normal form. So this table indicates that majority of Female patients whose age is 40-90 have no diabetes until they carries Normal tests. This data has 320 association rules.

The Graph of this table is shown as under.

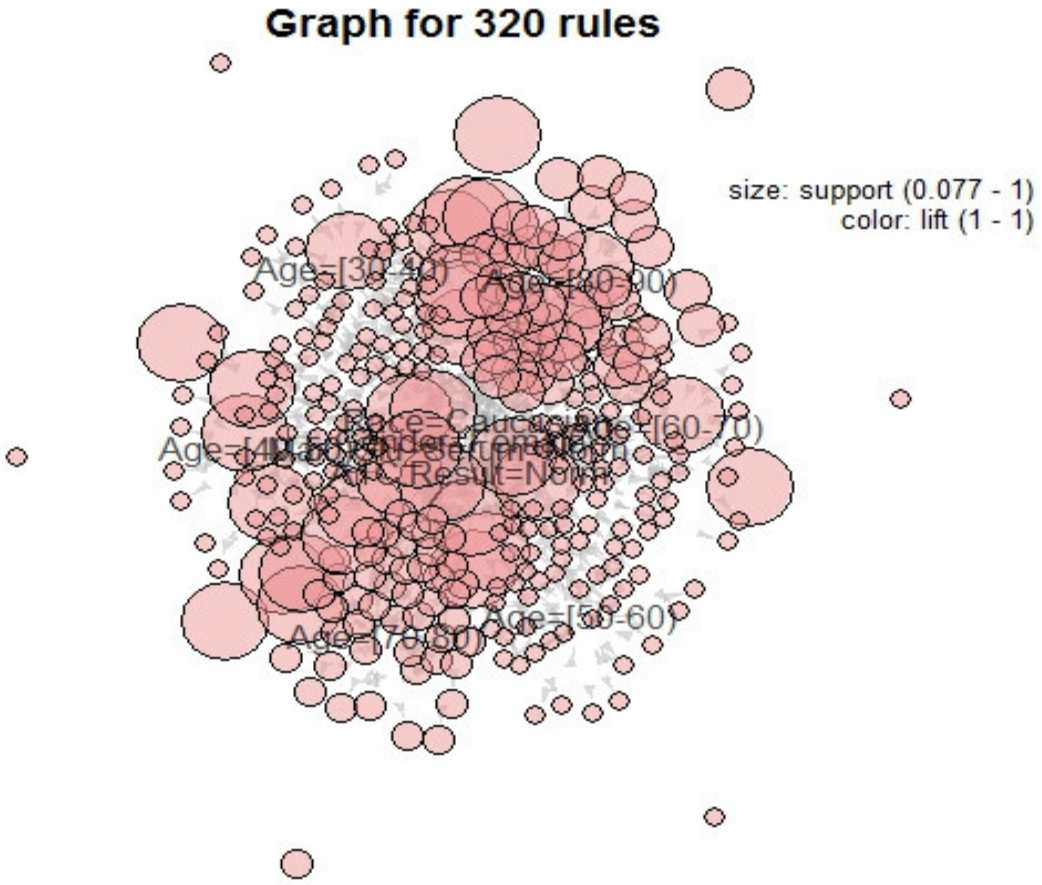


Here another table is presented for further illustration.

Race	A1C result	Max_glu_serum	Age	Gender
Caucasian	Norm	Norm	[70-80)	Male
Caucasian	Norm	Norm	[80-90)	Male
Caucasian	Norm	Norm	[40-50)	Male
Caucasian	Norm	Norm	[80-90)	Male
Caucasian	Norm	Norm	[70-80)	Male
Caucasian	Norm	Norm	[60-70)	Male
Caucasian	Norm	Norm	[60-70)	Male
Caucasian	Norm	Norm	[50-60)	Male
Caucasian	Norm	Norm	[70-80)	Male
Caucasian	Norm	Norm	[80-90)	Male
Caucasian	Norm	Norm	[40-50)	Male
Caucasian	Norm	Norm	[70-80)	Male

Table: 6 Caucasian (Normal) Male

This Table contains data of Caucasian race accomplished of Male gender who's both tests like A1C Result and Max_glu_serum are in Normal form. So this table indicates that majority of Male patients whose age is 40-90 have no diabetes until they carries Normal tests. This data has 320 association rules.
The Graph of this table is shown as under.



Likewise this 17 tables has been generated from the dataset according to Tests result combination separated on gender base. Male and Female table are separately created. All the tables have been processed by using R language. Association rules are calculated by using R language. Graphs has been created for the analysis of our research.