International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

# Risk Analysis of Diabetes Mellitus by Association Rule Summarization

**Aishwarya Ingle[1], Shivani Paraskar[1], Avanti Ajane[1], Sneha Mohadikar[1], Lokesh Thota[1], Prof. Ashwini Yerlekar[2]**

[1]BE Scholar, Department of Computer Science and Engineering Rajiv Gandhi College of Engineering and Research, Nagpur, Maharashtra, India

[2]Assistant Professor, Department of Computer Science and Engineering Rajiv Gandhi College of Engineering and Research, Nagpur, Maharashtra, India

## ABSTRACT

Early detection of patients with elevated risk of diabetes is very important in order that patients will begin to manage diabetes early and probably stop or delay the intense disease complications. By applying association rule mining to Electronic Medical Records (EMR), we intend to discover the set of risk factors and their respective collection that betokens the patients at particularly high risk of enrooting diabetes. We studied three association rule summarization technique and did a relative evaluation of these methodologies. We made use of these methodologies to find the fundamental segments which incite high risk of diabetes. All these three strategies made summations that portrayed sub masses at a high threat of diabetes with each system having its unmistakable quality. According to our inspiration, we use bottom up summarization (BUS) which conveys more fitting summary.

**Keywords**: Diabetes Mellitus, Data mining, Association Rule Mining, Survival Analysis, Association Rule Summarization

## I. INTRODUCTION

Diabetes is a crucial public pathological state and a non-contagion targeted for action by world leaders. Both the number of cases and the pervasiveness of diabetes have been linearly increasing over the past few decades. Globally, calculable 422 million adults were suffering from diabetes in 2014, compared to 108 million in 1980. The global pervasiveness (age-standardized) of diabetes has nearly doubled since 1980, linearly increasing from 4.7% to 8.5% in the adult population. This reflects an increase in associated risk factors such as being overweight or obese. Over the past 10 years, diabetes pervasiveness has risen rapidly. Diabetes of all types (Type I, Type II) can lead to complications in many parts of the body and can increase the overall risk of dying at an early age. Possible complications include ischemic heart disease, stroke, nephropathy, retinopathy, neuropathy and peripheral vascular disease. In pregnancy, poorly controlled diabetes increases the risk of fatal death and other complications.

Diabetes is a chronic disease, which is a constellation of diseases including hyperlipidaemia, hypertension, and central obesity. These diseases interact with each other, with cardiac and vascular diseases and thus studying and casting these interactions is important. We are making use of association rule mining so as to associate potentially interacting diseases with the elevated risk. The usage of this technology is favourable because it is not only evaluating the risk but also providing the physician with a justification, thus providing associated set of conditions. In this, work we are making use of rich risk factors. These risk factors include co morbid diseases, laboratory results, and medications. This information is easily available with electronic medical records (EMR) systems. With such a vast set of risk factors the generated rules grows to a size that critically cripples the interpretation. Thus for easy interpretation of the rules, we are compressing them by making use of association rule set summarization technique. The main goal of this palimpsest is to compare and characterize the three association rule summarization technique

Diabetes is a touch of the metabolic issue, which is an awesome grouping of diseases including hyperlipidaemia (lifted triglyceride and low HDL levels), (hypertension) and focal weight (with weight record beating 30 kg/m2). These maladies talk with each other, with cardiovascular and vascular ailments and along these lines recognition and demonstrating these associations is fundamental. Partnership standards are proposition that associate a strategy of conceivably cooperating conditions (e.g. high BMI and the closeness of hypertension analysis) with lifted hazard. The utilization of association benchmarks is especially great, in light of the way that in spite of evaluating the diabetes shot, they likewise right away give the pro a "protect", especially the related strategy of conditions. This strategy of conditions can be utilized to control treatment towards a more revamp and focused on preventive care or diabetes association. While association rules themselves can be enough deciphered, the subsequent supervise sets would some have the capacity to of the time be significant, disintegrating translate furthest reaches of the lead set when all is said in done.

Especially, in this work, we consider a rich game plan of risk segments, specifically co-dreary infirmities, lab results, medications and measurement data that are normally open in electronic remedial record (EMR) systems. With such a wide plan of risk factors, the course of action of discovered norms ends up noticeably combinatorial far reaching, to a size that greatly forestalls interpretation. To beat this test, we associated oversee set summarization techniques to pack the principal lead set into a more traditionalist set that can be interpreted easily. Different viable association represent set abstract methodologies have been proposed [10] yet no unmistakable course exists regarding the genuine nature, characteristics and weaknesses of these methodology. The centralization of this unique duplicate is to review and portray four existing association lead rundown frameworks and provide guidance to experts in picking the most sensible one. A run of the mill inadequacy of these frameworks is their inability to consider diabetes risk– a constant outcome–. Remembering the ultimate objective to make these strategies more appropriate, we expected to irrelevantly change them: we extend them to merge data about endless outcome factors.

Specifically, our key responsibilities are according to the accompanying.

- ✓ We present a clinical use of alliance lead mining to recognize sets of co-disheartening conditions (and the patient sub peoples who encounter the evil impacts of these conditions) that recommend out and out extended peril of diabetes.
- ✓ Association control mining on this wide game plan of components achieved an exponentially far reaching course of action of association measures. We created four common connection run set outline frameworks (fundamentally from the review [10]) by combining the threat of diabetes into the path toward finding a perfect once-over.
- ✓ Our essential responsibility is a relative evaluation of these enhanced framework frameworks that provides guidance to specialists in picking a fitting figuring for a practically identical issue.

## II. RELATED WORKS

A diabetes list is in a general sense an insightful model that apportions a score to a patient in perspective of his assessed peril of diabetes. Collins et al. [7] coordinated a wide audit of diabetes records portraying the peril factors and the showing framework that these documents utilized. They found that most records were included substance in nature and none of the outlined documents have thought about collaboration among the peril components.

While we don't think about any new diabetes document conveyed after the diagram, a present survey [12] focusing on the metabolic issue (of which diabetes is a section) addresses a basic headway. Kim et al. used alliance run mining to purposely examine co-events of conclusion codes. The consequent alliance rules don't constitute a diabetes record in light of the fact that the survey does not appoint a particular aftereffect of interest and they don't assess or suspect the threat of diabetes in patients, be that as it may they discovered some enormous connection between discovering codes.

We have starting late endeavoured a diabetes consider [4] where we anticipated that would discover the associations among afflictions in the metabolic issue. We used an unclear friend from this present survey; regardless, we included only eight discovering codes and age as indicators. We discovered alliance rules including some of these eight discovering codes, reviewed the peril of diabetes that these standards give on patients and presented the principles as a development chart outlining how patients progress from a strong state towards diabetes. We showed that the approach found clinically vital association concludes that are consistent with our helpful want.

With only eight pointer factors, the degree of the discovered oversee set was modest – 13 important rules– and along these lines, explanation was immediate. As a matter of fact, no administer set rundown was central.

## III. LITERATURE SURVEY

Chaudhari et al [13] Disease assertion is a champion among the most basic employments of such structure as it is one of the rule wellsprings of going wherever all through the world. Anticipate the human use the obligations from complex tests encouraged in labs additionally expect the disease considering danger parts, for instance, tobacco smoking, alcohol assertion, age, family history, diabetes, hypertension, raised cholesterol, physical laziness, weight. Specialists have been using a few data mining methods to help medicinal associations' experts in the examination of coronary disease. K-Nearest-Neighbor (KNN) is one of the plausible data mining methodology used as a touch of interest issues. Starting late, specialists are showing that joining particular classifiers through voting is conquering other single classifiers. This paper gets some information about applying KNN to help human association's aces in the total of infirmity exceptionally coronary sickness. It likewise inquires as to whether masterminding voting with KNN can update its precision in the confirmation of coronary affliction patients. The results show that applying KNN could achieve higher precision than neural framework assembling in the finding of coronary ailment patients. The results in addition exhibit that applying voting couldn't redesign the KNN exactness in the affirmation of coronary disease.

Prof. Mythili et al [12] Diabetes mellitus, in key terms called as diabetes, is a metabolic sickness, where a man is affected with high blood glucose level. Diabetes is a metabolic issue sped up in light of the foul up of body to make insulin or to properly utilize insulin. This condition rises when the body does not make enough insulin, or in light of the way that the cells don't respond to the insulin that is passed on. Blood glucose test is the fundamental system for diagnosing diabetes. Moreover, there have been assorted robotized methodology proposed for finish of diabetes.

Each and every one of these frameworks has a couple of data regards which would be the delayed consequence of different tests that should be done in recuperating center interests. This paper proposes a structure those strategies to help the patients encountering arranged restorative tests, which by far most of them consider as a dull undertaking and excess.

The parameters saw for diagnosing diabetes have been made in a way that, the customer can expect in case he is influenced with diabetes himself. Back Propagation count is used for conclusion.

Ahmed et al [15] Heart affliction is an essential explanation for bleakness and mortality in cutting edge society. Medicinal conclusion is basic yet got undertaking that should be performed unequivocally and enough. The capable data examination instruments are used to remove satisfying getting from the immense measure of supportive data. There is huge data open inside the medicinal associations' structures. In any case, there is an endeavour of influencing examination gadgets to discover campaigned associations and cases in data. Learning presentation and data mining have found diverse applications in business and trial space.

One of the applications is contamination finding where data mining instruments are indicating helpful outcomes. This examination paper proposed to find the heart issue through data mining, Support Vector Machine (SVM), Genetic Algorithm, repulsive set theory, association rules and Neural Networks. In this study, we speedily assessed that out of the above systems Decision tree and SVM is best for the coronary issue. So it is watched that, the data mining could help in the seeing check or the need of high or everything considered safe heart issue.

Thangaraju et al [16] Data mining is the appearing of looking earlier databases with a particular real objective to make new data. There are particular sorts of data mining strategies are available. Procedure, Clustering, Association Rule and Neural Network are likely the hugest structures in data mining. In Health mind associations, Data mining recognize an essential part. Most an unbelievable bit of the time the data mining is used as a touch of human associations tries for the course toward foreseeing diseases. Diabetes is an unending condition. This proposes is continues for a long time, dependably for some individual's whole life [11]. This paper considers the examination of diabetes gaging approaches using gathering structures. Here we are using three one of a kind sorts of amassing frameworks named as Hierarchical party; Density based collecting, and Simple K-Means gathering. Weka is used as a gadget.

Durairaj et al [17] Neural Networks are one of the delicate enlisting frameworks that can be used to make needs on healing data. Neural Networks are known as the Universal pointers. Diabetes mellitus or essentially diabetes is a contamination achieved in light of the increase level of blood glucose.

Specific standard system, considering physical and advancement tests, are open for diagnosing diabetes. The Artificial Neural Networks (ANNs) based system can satisfactorily relate for hypertension chance need. This upgraded display disconnects the dataset into both of the two parties. The earlier revelation using delicate enrolling systems help the specialists to lessen the probability of ceasing all clowning around of the disease. The data set chose for request and exploratory energy relies upon Pima Indian Diabetic Set from (UCI) Repository of Machine Learning databases. In this paper, a confined audit is driven on the utilization of different sensitive arranging structures for the need of diabetes. This outline is relied on to see and propose a beneficial technique for earlier check of the infirmity.

## IV. IMPLEMENTATION

We endeavor to use association rule mining to the electronic helpful record (EMR); All the risk ascertain about a patient specifically co-bleak illness and research office results and arrangements are being available in the EMR, there are less chances to miss experiences about a patient with the expansive game plan of danger factors the course of action of observed peril ends up being incredibly huge to vanquish this we use rule set layout strategy which is used to pack the main rule set into a traditionalist set. We use the going with frameworks:

- ✓ RPG-global
- ✓ TOPK
- ✓ BUS.

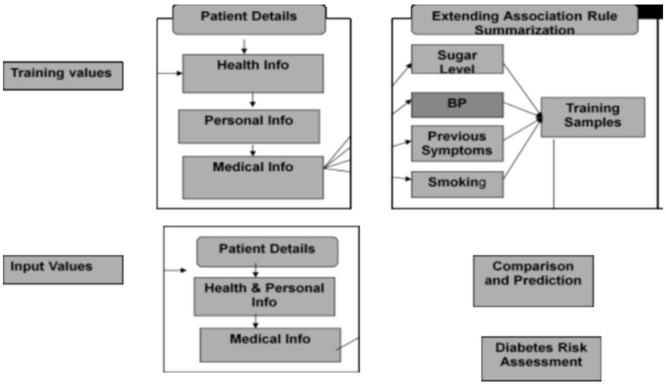

**Figure 1.** System Architecture

We first check the help of individual things and make sense of which of them broad (ie.) we have minimum help are. In each pass we start with a seed set of things saw to be sweeping in the past pass. We use this seed set for making new potentially generous thing sets called contender key, thing set and number the bonafide bolster for these candidate thing set in the midst of the nonchalance the data.

At the complete of the pass we make sense of which confident thing set are actually broad and they advance toward getting to be seed for not pass. This methodology contains until the point that no new considerable thing set are found. Testing quantifiable criticalness: for each discovered thing set we have to test whether the outcome scattering in the impacted and unaffected subpopulation is unmistakably particular.

Step-2 the course of action of thing set is isolated with the goal that solitary the verifiably basic ones are returned as distributional association rule, this rule is depicted by the going with bits of knowledge from the amount of thing set assembled. Let OR be the watched number of diabetes scene in the subpopulation DR secured by R. allow ER to mean the typical number of diabetes events in the subpopulation secured by R.

$ER = OR\text{-}i\varepsilon DRyi$ where yi is the martingale for patient.

The relative risk factor is defined by R

$RR = OR/ER.$

**Table 1.** Description of the risk factors that appeared in any of the summarized rules

| Parameter | Weightage | Values |
|---|---|---|
| Male & Female | Age<30 >30to<50 | 0.1 0.3 0.7 0.8 |
| Smoking | Never Past Current | 0.1 0.3 0.6 |
| Overweight | Yes No | 0.8 0.1 |
| Alcohol intake | Never Past Current | 0.1 0.3 0.6 |
| Heart rate | Low(<60 bpm) Normal(60 to 100bpm) High(>100bpm) | 0.9 0.1 0.9 |
| Blood sugar | High(>120&<400) Normal(>90&<120 ) Low(<90) | 0.5 0.1 0.4 |
| Bad cholesterol | Very high>200 High(160 to 200) Normal<160 | 0.9 0.8 0.1 |

When we endeavor to apply distributional rule mining with our electronic therapeutic records it made a far reaching number of (quantifiably basic) rules. Rules that were created possibly shift from each other provoking obfuscating of clinical patters. With a particular true objective to vanquish the issue of this broad number of rules which were delivered we go for condensing the rule set into more diminutive set for our less requesting audit. We first study the present rule set and database diagram systems then we endeavor to meld a nonexclusive structure with a particular true objective to get a tenacious consequence of variable into record.

By and by we demonstrate the rule set delivered by the extended layout calculation, for each calculation we used the parameter setting that gave the best results to APRX Collection we used α = 0.1, λ = 1 for RPG overall we used δ = 0.5, σ = 0.2, λ = 0.98 for top K we used λ = 0.2 and for BUS we used λ = 0.1

## A. RPG-global

The rule drawbacks of some algorithms like APRX Collection were the abundance in the rule set and the debilitating of the danger. The RP-Global summary resembles APRX Collection n in that it is generally stressed with the surge of the rule and subsequently it plays out a to a great degree intense weight. RPG Global has two drawbacks by thinking about Patient extension and by building the layout from rules in the primary rule set.

**Table 2.** Rule set created by RPGlobal.

| RR | ER | OR | RULE |
|---|---|---|---|
| 1.69 | 32 | 55 | Bmi trigal acearb diuret htn |
| 1.23 | 52 | 65 | Acearb bb diuret aspirin htn |
| 1.29 | 42 | 55 | Sbp tchol acearb diuret htn |
| 2.10 | 25 | 54 | Hdl trigal diuret aspirin htn |
| 1.28 | 42 | 54 | Bmi tchol hdl trigl tobacco |

## B. TOPK

Top-K algorithm diminishes the reiteration in the rule set which was possible through taking a shot at patients rather than the assertion of the rules. This approach surrendered the exceptional weight rates of past two calculation TOP-K still achieves high weight rate and it adequately recognized rules with high danger and low abundance.

**Table 3.** Rule created by the top-k algorithm

| RR | ER | OR | RULE |
|---|---|---|---|
| 2.40 | 21.70 | 52 | Fibra htn |
| 1.58 | 37.97 | 60 | Bmi hdl ihd |
| 1.47 | 45.52 | 67 | Sbp htn tobacoo |
| 1.46 | 317.03 | 464 | Bmi htn |
| 1.62 | 32.16 | 52 | Sbp tchol trigal statin htn |

## C. BUS

The layouts made by BUS (showed up in Table 5) and Top-K are tantamount in quality. The BUS diagram demonstrates less variability (it has a tendency to use comparable conditions: bmi and trigl co-occur in 40% of the rules), yet this reduced vacillation does not change over into extended abundance in the patient space.

Transport (as opposed to Top-K) deals with the patients and not on the rules. In like manner, overabundance to the extent rule articulation can happen. In any case, BUS unequivocally controls the abundance in the patient space through the parameter charging the base number of new (as of now uncovered) cases (patients with diabetes event) that ought to be secured by each rule. In this way the reduced variability in the rule articulation does not change over into extended reiteration.

**Table 4.** Top 10 summarized rule created by BUS

| RR | ER | OR | RULE |
|---|---|---|---|
| 2.34 | 24 | 57 | Bmi trigal acearb statin htn |
| 2.10 | 25 | 54 | Hdl trigal diuret aspirin htn |
| 1.91 | 56 | 107 | Bmi trigal statin htn |
| 1.54 | 78 | 121 | Bmi trigal tobacco |
| 1.37 | 39 | 54 | Dbp diuret htn |

The BUS rule set made sense of how to fuse coronary sickness earlier (rule #3) and with higher peril (2.15) than the Top-K rule set. In like manner, BUS includes tobacco use in a mix of danger factors with higher relative risk than Top-K. As a rule, paying little mind to the refinements in the calculations, BUS and Top-K create practically identical phenomenal once-overs.

## V. EXPERIMENTAL RESULT

The measure of rules should have been lessened to a level where clinical elucidation is plausible. To this end, we centered four strategies to pack these rules into sets of 10-20 rules that clinical experts can assess.

While every single one of the four methodology made sensible rundowns, every framework had its unmistakable quality. In any case, not these attributes are fundamentally beneficial to our application. We found that the most fundamental differentiator between the computations is whether they utilize an affirmation model to join a rule in the structure in context of the surge of the rule or in light of the patient sub masses that the rule covers.

APRX-COLLECTION and RPGlobal in a general sense wear down the surge of the rules with an essential focal point of broadening weight. They utilize delegate rules, each of which tends to various uncommon rules. Such illustrative rules accomplish high weight, however weaken the danger of diabetes over the as often as possible massive subpopulation they cover. Top-K and BUS work in a general sense on the patients and their objective– especially if there should be an occurrence of Top-K– can be considered as confining excess. They passed on amazing designs in light of the way that a strong reaction of diminishing repetition is to accomplish mind blowing weight. The reverse isn't genuine: high weight rate does not understand low emphasis.

Among Top-K and BUS, we found that BUS held conceivably more repetition than Top-K, which enabled it to have better patient degree and better capacity to patch up the primary data base. This perfect position made BUS the most appropriate count for our motivation.

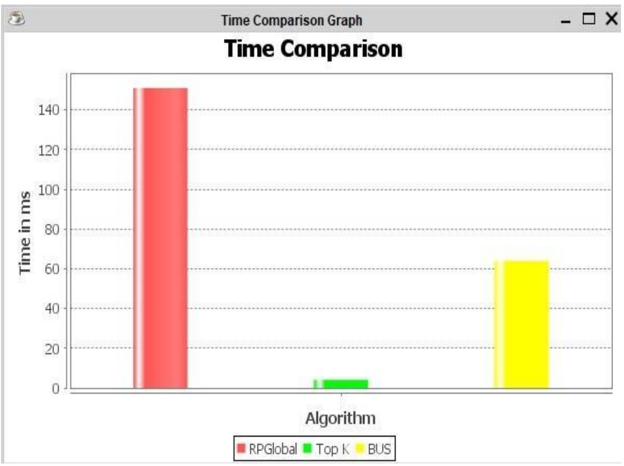

**Figure 2.** Time Comparison

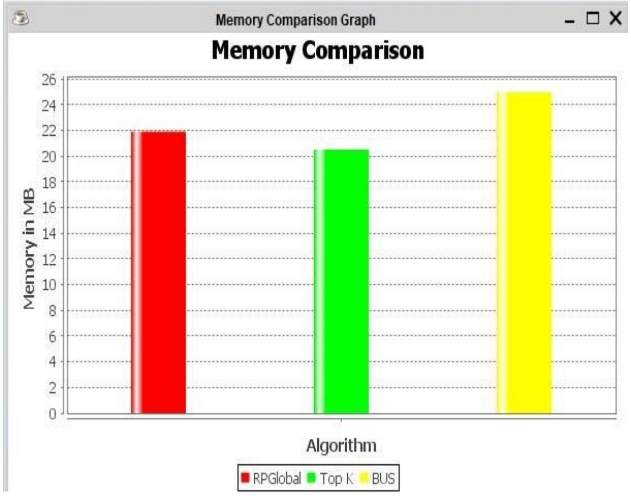

**Figure 3.** Memory Comparison

Our Result also shows the Time and space utilization of the above four algorithms. Figure show the results generated for the same.

## VI. CONCLUSIONS

The electronic data made by the utilization of EMRs in routine clinical practice can bolster the divulgence of new learning. Cooperation control mining coupled to a unique system gives an essential contraption to clinical research. It can reveal secured clinical associations and can propose new instances of conditions to divert killing movement, association, and treatment approaches. In our particular portrayal, we utilized distributional association control mining to perceive sets of hazard segments and the relating understanding sub populaces that are at totally stretched out danger of advancing to diabetes. An over the best amounts of association rules were found impeding the clinical clarification of the outcomes.

## VII. REFERENCES

[1]. F. Afrati, A. Gionis, and H. Mannila, "Approximating a collection of frequent sets," in Proc. ACM Int. Conf. KDD, Washington, DC, USA, 2004.

[2]. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th VLDB, Santiago, Chile, 1994.

[3]. Y. Aumann and Y. Lindell, "A statistical theory for quantitative association rules," in Proc. 5th KDD, New York, NY, USA, 1999.

[4]. P. J. Caraballo, M. R. Castro, S. S. Cha, P. W. Li, and G. J. Simon, "Use of association rule mining to assess diabetes risk in patients with impared fasting glucose," in Proc. AMIA Annu. Symp., 2011.

[5]. Centers for Disease Control and Prevention. "National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States," U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011 Online].

[6]. V. Chandola and V. Kumar, "Summarization-Compressing data into an informative representation," Knowl. Inform. Syst., vol. 12, no. 3, pp. 355–378, 2006.

[7]. 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," BMC Med., 9:103, Sept. 2011.

[8]. Diabetes Prevention Program Research Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," N. Engl. J. Med., vol. 346, no. 6, pp. 393–403, Feb. 2002.

[9]. G. Fang et al., "High-order SNP combinations associated with complex diseases: Efficient discovery, statistical power and functional interactions," PLoS ONE, vol. 7, no. 4, Article e33531, 2012.

[10]. M. A. Hasan, "Summarization in pattern mining," in Encyclopedia of Data Warehousing and Mining, 2nd ed. Hershey, PA, USA: Information Science Reference, 2008.

[11]. Rakesh Motka, Viral Parmar, Balbindra Kumar, A. R. Verma, " Diabetes Mellitus Forecast Using Different Data Mining Techniques", International conference on computer and Communication Technology

[12]. Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, "Diagnosis of Diabetes Mellitus based on Risk Factors", International Journal of Computer Applications, Vol.10, Issue No.4, November.2010

[13]. Anand A. Chaudhari, Prof.S.P.Akarte, " Fuzzy and Data Mining based Disease Predection using K-NN Algorithm", International Journal of Innovations in Engineering and Technology, Vol. 3, Issue No. 4, April 2014

[14]. Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, " Diagnosis of Diabetes Mellitus based on Risk

Factors", International Journal of Computer Applications, Vol. 10, Issue No. 4, November 2010

[15]. Aqueel Ahmed, Shaikh Abdul Hannan, " Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue No. 4, September 2012

[16]. P. Thangaraju, B.Deepa, T.Karthikeyan, "Comparison of Data mining Techniques for Forecasting Diabetes Mellitus", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue No. 8, August 2014

[17]. M. Durairaj, G. Kalaiselvi, " Prediction Of Diabetes Using Soft Computing Techniques- A Survey", International Journal of Scientific & Technology Research, Vol. 4, Issue No.3, March 2015

[18]. S.F.B, Jaafar and Darmawaty Mohd Ali. "Diabetes Mellitus Forecast using Artificial Neural Network (ANN), Asian conference on sensors and the international conference on new techniques in pharmaceutical and medical research proceedings (IEEE), Kuala Lumpur, Malaysia, 5-7 September 2005, pp 135-139.

[19]. S. Alby, B. L. Shivakumar," A survey on data-mining technologies for prediction and diagnosis of diabetes", International conference of IEEE 2014.

[20]. Gyorgy J. Simon, Terry M. Therneau, Steven S. Cha, " Extending association rule summarization techniques to assess risk of diabetes mellitus", IEEE VOL 27, no. 1, January 2015