Knowledge Discovery in Health Care Datasets Using Data Mining Tools

MD. Ezaz Ahmed
Department of Computer Science & Engineering
ITM University,
Gurgaon, India

Dr. Y.K. Mathur
183 First Floor, Vaishali, Delhi University
Teacher's Housing Society
Delhi, India

Dr Varun Kumar
Head of Department
Department of CSE
MVN,
Palwal, India

***Abstract*—Non communicable diseases (NCDs) are the biggest global killers today. Sixty-three percent of all deaths in 2008 – 36 million people – were caused by NCDs. Nearly 80% of these deaths occurred in low- and middle-income countries, where the highest proportion of deaths under the age of 70 from NCDs occur [1]. The commonness of NCDs, and the resulting number of related deaths, are expected to increase substantially in the future, particularly in low and middle-income countries, due to population growth and ageing, in conjunction with economic transition and resulting changes in behavioral, occupational and environmental risk factors. NCDs already disproportionately affect low and middle-income countries. Current projections indicate that by 2020 the largest increases in NCD mortality will occur in Africa, India and other low and middle-income countries [2].**

Computer-based support in healthcare is becoming ever more important. No other domain has so many innovative changes that have such a high social impact. There has already been a long standing tradition for computer-based decision support, dealing with complex problems in medicine such as diagnosing disease, managerial decisions and assisting in the prescription of appropriate treatment. As we know that "Research is for the people not for yourself" so we are pleased to work for the healthcare and hence for the society and ultimately the MANKIND.

Keywords: NCDs; Web; Web Data; Web Mining; data Mining Healthcare.

I. INTRODUCTION

Healthcare researchers as well as practitioners require a lot of information to make their healthcare related activities and practices either with drug prescriptions which can efficiently cure patients' illness or with correct and efficient medical/clinical procedures and services. Over the last decade, we have witnessed a likely to explode growth in the information available on the World Wide Web. Today, web browsers provide easy access to innumerable sources of text and multimedia data. More than 1000000000 pages are indexed by search engines, and finding the preferred information is not an easy task. This abundance of resources has provoked the need for developing automatic mining techniques on the World Wide Web, thereby giving rise to the term "web mining." Information technology has been playing a vital and critical role in this field for many years. Therefore being able to promptly and correctly access required medical databases and information system resources and effectively communicate across different medical Institutes or countries become necessary. To proceed toward web intelligence, requires the need for human intrusion, we need to integrate and embed knowledge discovery, and machine learning into web tools.

The Web has become a major vehicle in performing research and practice related activities for healthcare researchers and practitioners, because it has so many resources and potentials to offer in their specialized professional fields. There is tremendous amount of information and knowledge existing on the Web and waiting to be discovered, shared and utilized. The research in improving the quality of life through the Web has become attractive. This paper summarizes. The reason for considering web mining, a separate field from data mining, is explained. The limitations of some of the existing web mining methods and tools are enunciated, and the significance of proposed model in health care. Scope for future research in developing "Proposed model of web mining" systems is explained. We present an approach regarding Semantic Web and mining [3] in healthcare, which can be used to not only improve the quality of Web mining results but also enhances the functions and services and the interoperability of medical information systems and standards in the healthcare field.

The objective of this article is to present an outline of web mining, Knowledge web mining, its subtasks, and to give a perspective to the research community about the potential of applying knowledge discovery techniques to its different components. The article, gives emphasis on possible enhancements of these tools using "Knowledge Web Mining". It should be noted that the use of knowledge discovery in "web mining" is a field in its origins, and thus the worth of this paper at this point in time is evident.

The rest of this paper is organized as follows: Section II deals with the web and web mining, knowledge web mining discussed in next Section i.e. in section III. Section IV provides an introduction to knowledge discovery. Sections V cover, in detail, the possible healthcare application and future model a practical approach, Section VI provides the conclusion and scope of future research in the area of knowledge web mining by proposed model.

II. WEB AND WEB MINING

Web is a collection of inter-related files on one or more Web servers. The web is a immense collection of completely uncontrolled heterogeneous documents.

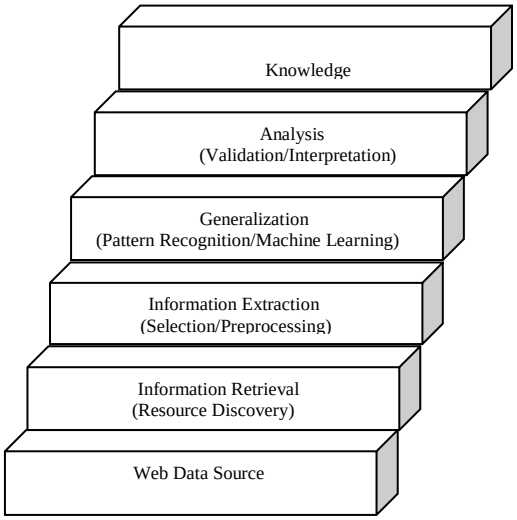

Figure1. The steps of extracting knowledge from data

Thus, it is huge, varied, and dynamic, and raises the issues of scalability, heterogeneity, and dynamism, respectively. Due to these characteristics, we are currently drowning in information, but famished for knowledge; thereby making the web a fertile area of data mining research with the vast amount of information available online. Data mining refers to the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Web mining can be mostly defined as the discovery and analysis of useful information from the World Wide Web.

Web mining is the application of data mining techniques to Web data [5]. Web mining helps to solve the problem of discovering how users are using Web sites. It involves mining logs (or log analysis) and the steps that typically have to be gone through to get meaningful data from Web logs - data collection, pre-processing, data enrichment and pattern analysis and discovery as given in figure 1.

Web mining is the application of data mining techniques to extract knowledge from Web data

Web data is

- Web content –text, image, records, etc.
- Web structure –hyperlinks, tags, etc.
- Web usage –http logs, app server logs, etc.

In web mining data can be collected at the server side, client side, proxy servers, or obtained from an organization's database. Depending on the location of the source, the type of collected data differs. It also has extreme variation both in its content (e.g., text, image, audio, symbolic etc.) and meta information, that might be available. This makes the techniques to be used for a particular task in web mining widely varying.

Some of the issues which have come to light, as a result, concern

- Need for handling context sensitive and imprecise
- Queries;
- Need for summarization and deduction;
- Need for personalization and learning.

Thus, web mining, though considered to be a particular application of data mining, warrants a separate field of research, mainly because of the aforesaid characteristics of the data and human related issues.

III. KNOWLEDGE WEB MINING

Relevance of knowledge web mining is extensively established in the literature recently, the application of knowledge web mining in health care problems has also drawn the attention of researchers.

Thus better healthcare related recommendations can be constructed with ontologies and with little human intervention. For an on-line healthcare web site, two important ontologies would need to be built: one of the ontology describing all the healthcare services provided, with the relation between each other, and the other ontology describing the web site. Thus Semantic Web ontology can help build better web mining analysis in healthcare, and web mining in-turn helps construct better, more powerful ontology in healthcare. Web personalization is to display and offer information to the healthcare web site users according to their interests and needs, which are already stored in the database. Personalization requires implicitly or explicitly collecting web site users information and leveraging that knowledge in the content delivery framework to choose what information to present to the users and how to present it with tailored pages according to information gathered about the particular health care web site user. Web mining is the application of data mining techniques to Web data.

Web mining helps to solve the problem of discovering how users are using Web sites. It involves mining logs (or log analysis) and the steps that typically have to be gone through to get meaningful data from Web logs - data collection, pre-processing, data enrichment and pattern analysis and discovery. We have proposed a new type of intelligent model in health care which is web based, In which we use web mining and data mining.

IV. KNOWLEDGE DISCOVERY

A. Proposed Model

Our work is for the architecture of a web based decision support system model. Means we have to work on basically three areas.

a) Web based Model
b) DSS
c) Intelligent web based Model (IDSS)

This will give us a Web based DSS model [6] which is Intelligent. Our work area will be health care, so we need data related to health care such as cardio, OSA, diabetes, breast

cancer etc. Now for the development of the model we have to use Data Mining Tools such as Weka, Tanagra and PSW @12 Modeler, Statistica.

As per our research related literature survey we came to know what was the traditional model what is the current model and what model we are going to proposed. All this will be discussed below and we have to work accordingly.

V. POSSIBLE HEALTHCARE APPLICATION AND FUTURE MODEL A PRACTICAL APPROACH

Now finally we are going to propose our model in which we have to use Web and Web mining. From web we fetch data and through web and web server we store data and after mining we make these data intelligent using our model. In this model we use data mining, web mining tools, web and web servers. Our proposed model will be as follows.

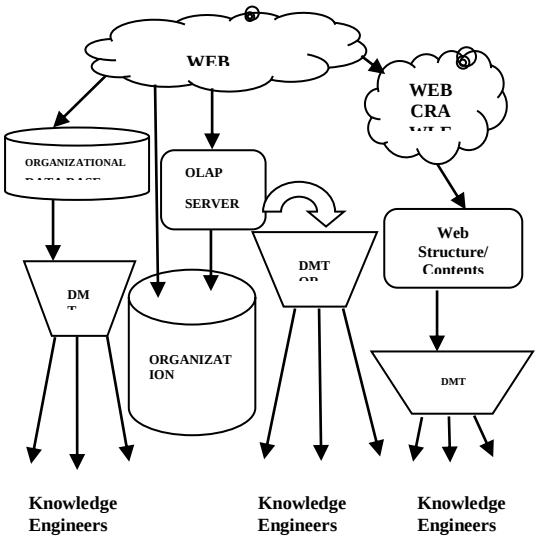

Figure2. Proposed Web based IDSS Model

As given in the above figure our research will base on three modes. In first mode we fetch data from the web and that organizational data base will be treated by data mining. By the help of data mining tools (DMT) or Knowledge mining Tools (KMT) we get knowledge data that will be used by the knowledge engineers or the users.

In second method our research will based on data which will be fetched by the web [6]. We fetch data and collect data as organizations' operational database. By the help of OLAP server we fetch data and on that data we apply DMT or KMT. This will result as knowledge data. Data will be further used by the knowledge engineers or the users. These data would be further send to web for the global use. Similarly in the third mode we fetch data from web by different web crawlers. In these data we have web structure, contents and web usage. On these data we use DMT or KMT and finally acquire knowledge data and that would be further used by knowledge engineers or end users given in figure 2.

A. *Diabetes and data mining*

When considering the healthcare business, we may find several interesting and demanding applications for DM [10]. Following our analytical formulation, we now present a real-life application for identifying diabetic patients in a small Indian town.

TABLE1: SHOWS THE PATIENT SUFFERING FROM THE DISEASE

Patient id	Disease 1	Disease 2	Disease 3
1	Blood Sugar	Blood Pressure	Heart Disease
2	Blood Sugar	Blood Pressure	Heart Disease
3	Blood Sugar	Blood Pressure	Heart Disease
4	Blood Sugar	Blood Pressure	Kidney Problem
5	Blood Sugar	Blood Pressure	Eye Problem

B. *Classification and Prediction*

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction [7].

IF-THEN rules are specified as IF condition THEN conclusion

e.g. IF age=old and patient=diabetic then heart disease prone=yes

C. *Clustering Analysis*

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity.

That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. [4]

Application of clustering in medical can help medical institutes' group individual patient into classes of similar behavior [7]. Partition the patient into clusters, so those patients within a cluster (e.g. healthy) are similar to each other while dissimilar to patient in other clusters (e.g. disease prone or Weak). As given in figure 3.

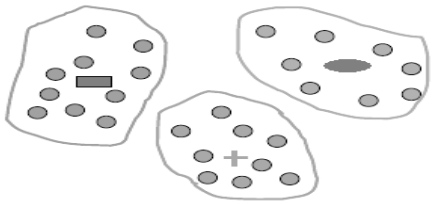

Figure 3. Picture showing the partition of patients in clusters

The main goal of this application is to be familiar with what causes diabetics. We were capable to obtain a patient database and conduct an analysis looking for to identify which patients have high probability of being diabetic.

Association Rules that can be derived from Table 1 are of the form:

$(X, disease1) \Rightarrow (X, disease2)$

$(X, disease1)\hat{}(X, disease2) \Rightarrow (X, disease3)$

$(X, "Bloodsugar") \Rightarrow (X, "Bloodpressure")$ [support=2% and confidence=60%]

(X,"Bloodsuger")^(X," Bloodpressure")
==>(X,"Heartdisease")
[Support=1% and confidence=50%]

Where support factor of the association rule shows that 1% of the patient suffering from the disease blood sugar and blood pressure, confidence factor shows that there is a chance that 50% of the patients who have "Blood sugar" will also have "Blood pressure".

This way we can find the strongly related disease and can optimize the database of a healthcare programme. As given in Table 1.

D. *Generating Knowledge*

The use of Information Theory is primarily interesting as this theory relates also to the Information Systems field. When integrates those concepts together we were capable to show that our method is relatively excellent compared to other traditional methods. Therefore, one outcome is establishing our method as a legitimate method for DM [8].

Second, we used to the DM procedure to gain knowledge about diabetes. We wrap up that the following variables can provide good indicators for identifying probable diabetic patients: family history, body weight (BMI Body Mass Index), pregnancy (in case of female patent), SFT (Skin Fold Thickness) and age. This may become a powerful predictive tool for any organization seeking to perform a more accurate and informed patient selection process to recognize diabetic patients.

E. *Working with Tanagra on Diabetes Dataset:*

Open Tanagra then open data file which is in txt, xls or arff format. We use tanagra for finding the class of particular disease and its various attributes such as Max, Min, Mean, std. deviation etc. then we find its class of diabetic and non diabetic patients then finally we are able to give description of both class of diabetic and non-diabetic patients [9] with there accuracy in percentage.

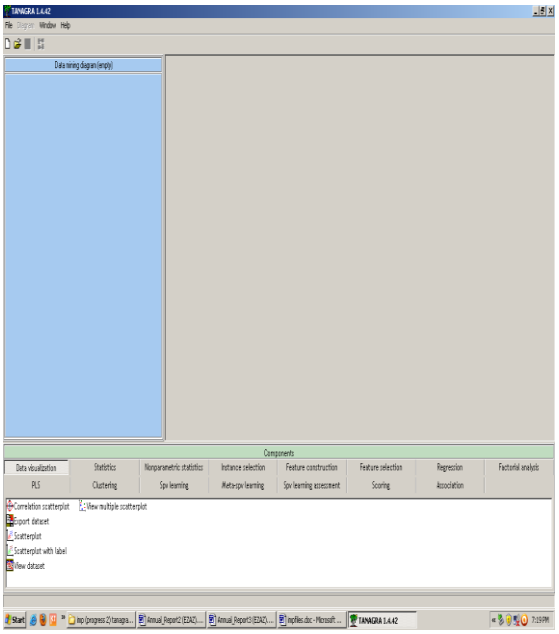

figure 4. Tanagra page

As we open the file in dataset it will appear as given below, the screen shot given below clearly indicate the open file name diabetic.txt in title diabetic class and also on the task bar of Tanagra. The down load information is on the right side of the screen given in figure 4 and 5.

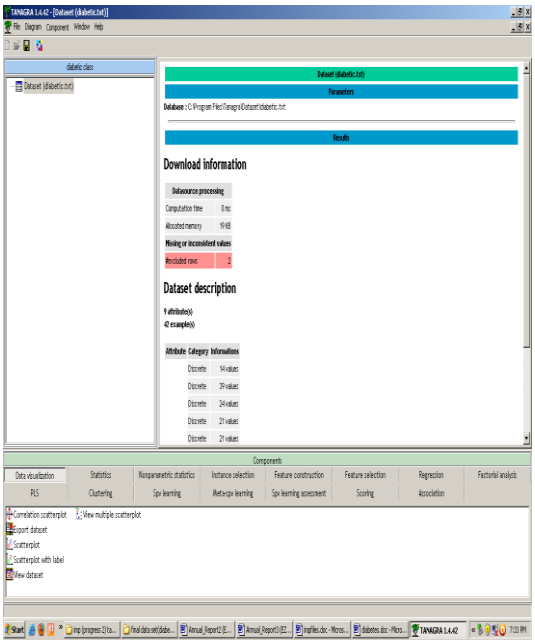

Figure 5. Open file information

First we open a new sheet of Tanagra and then open a dataset of diabetic patient in the format given in Tanagra. First look of Tanagra is given in figure 4. Dataset file name diabetic.txt. Now we right click on view dataset and choose view from pop-up menu which will appear after right click on view dataset. This gives the data on the Tanagra sheet given below in figure 6. Now we select view dataset from data

visualization tab and drag it and drop to that on dataset. Now we select define status from feature selection tab and then drag it and drop to dataset then right click on define status and select parameters from pop up menu we get above attributes given in figure 7.

Now select four attributes as input as age, BMI, DPF and Plasma Glucose and press OK button. Now we get following result given in figure number 8.

From statistics tab we choose Univariate continuous stat, drag and drop it in define status1. Then we use view command from pop up menu we will get following figure 8. In above figure it is clear that we get result as Min, Max, Average, Std-dev and avg. Std-dev.

From example Plasma Glucose min value is 78 max values is 197 Average 126.7 Std dev is 32.11 and avg. std dev is 0.253 BMI (Body Mass Index) result is 0, 45.8, 31.42, 7.9643, and 0.2534.

Similarly for age min age is 21 max ages in dataset is 60, average is 37.9024, std. dev is 11.6293 and avg. std dev is 0.3068 as given in above figure 8. Again we select define status3 and drag and drop group characterization figure 9 from statistics tab. Then press right click and choose view we will get following result in figure 9.

In the next section we have outcomes of the work explained.

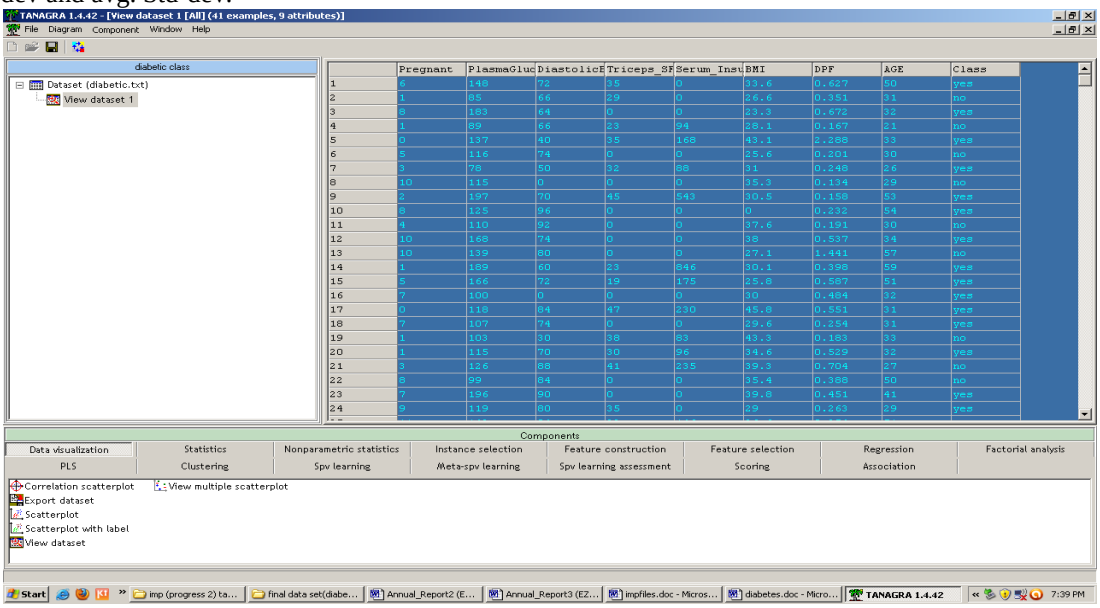

Figure 6. Diabetes dataset in Tanagra

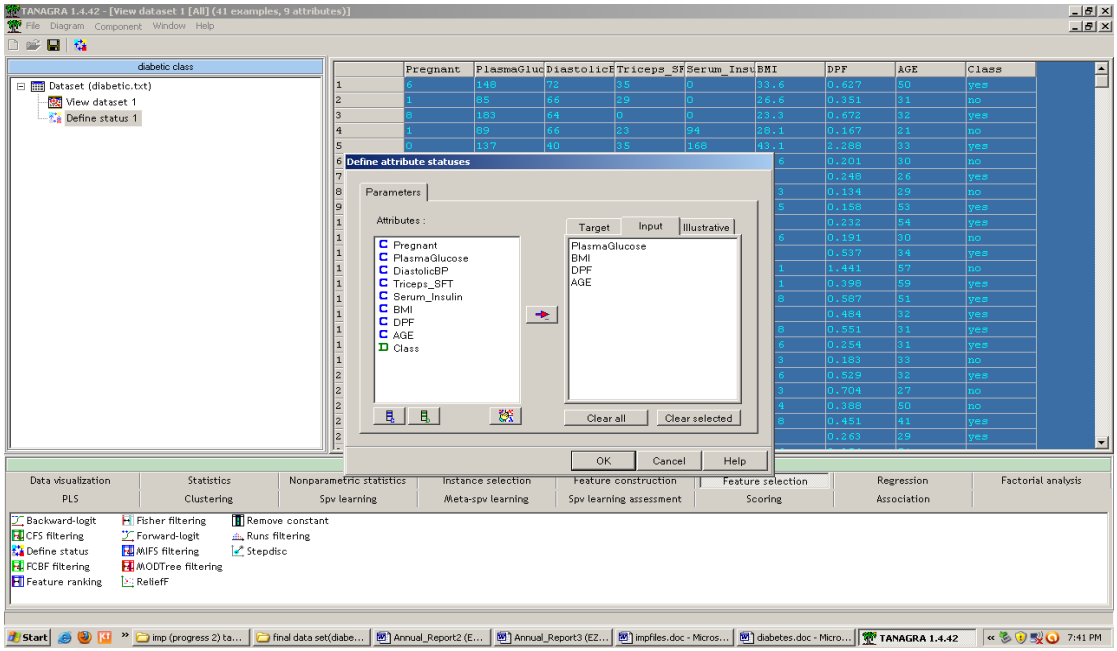

Figure 7. Selection of different attributes

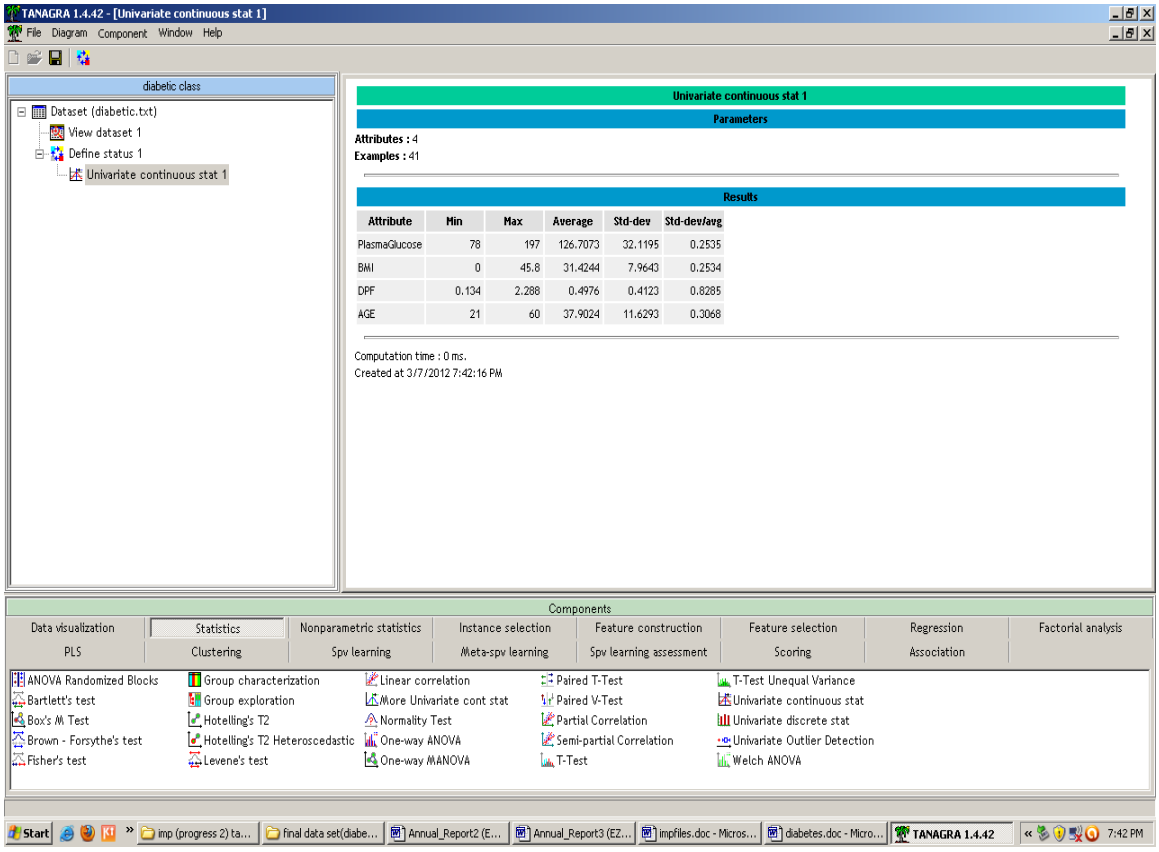

Figure 8. Min, Max, average, std. deviation of different attributes

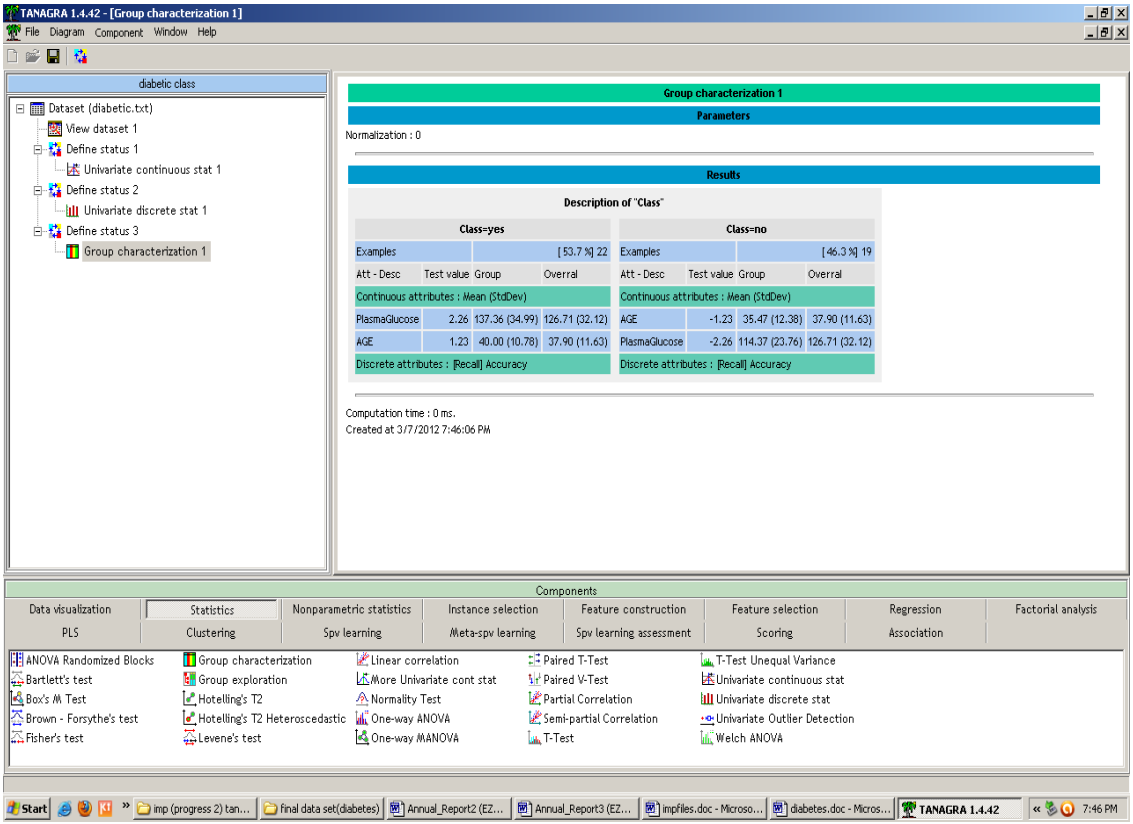

Figure 9. final outcomes on given diabetes dataset.

In this result we could conclude that the group 'Yes' means diabetic class having mean value of plasma glucose is 137.36 and standard deviation is 34.99 for the average mean age group is 40 years.

Whereas for group 'No' means non-diabetic class the mean plasma glucose value is 114.37 and standard deviation is 23.76 for average age group is 35.47. This is overall 41 patients clinical records test dataset. We could draw a conclusion that average age for non-diabetic is 35-36 years; increase of plasma glucose depends on after 40 years of age. So after 40 years of age a person is more prone to diabetic according to his/her plasma glucose value.

VI. CONCLUSION

As we continue our fight against diabetes, sharing and benchmarking diabetes care is essential to influence health policy and improve outcomes and quality of life for people with diabetes. As more and more data is collected, Diabetes measurement will become an even more powerful resource for inspiring and driving change in diabetes care. The fundamental goal of the Diabetes measurement is to measure, share, and improve diabetes outcomes. There are so many ways to get involved in reversing diabetes trends, from collecting data to sharing better practice models to improving public visibility and advocating for the quality of diabetes care at the global, national, clinic, and patient levels.

In this work we have presented an intelligent proposed model for healthcare which is related with healthcare. In this proposed model diabetic patient. This web based decision support system will helpful in health care management. Since the application of data mining brings a lot of advantages in higher well equipped hospitals, it is recommended to apply these techniques in the areas like optimization of resources, prediction of disease of a patient in the hospital.

As shown in the proposed model mentioned above, the main components of IDSS are intelligent techniques that generate knowledge which further helps health care planners to take more accurate decisions. Future work will be done on heart disease clinical dataset and find outcomes on heart disease. Work will be done on weka and will get knowledge data.

REFERENCES

[1] The world health report 2002: Reducing risks, promoting healthy life. Geneva, World Health Organization, 2002.

[2] Global health risks: mortality and burden of disease attributable to selected major risks. Geneva, World Health Organization, 2009.

[3] Gerd Stumme, Andreas Hotho, Bettina Berendt, "Usage Mining for and on the Semantic Web," *Proc. Of the Joint Conference on Information Sciences*, pp. 200-204, 2003.

[4] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," in Proc. 6th Int.WWWConf., 1997, pp. 391–404.

[5] B. Mobasher, N. Jain, E.-H. Han, and J. Srivastava, "Web Mining: Patterns from WWW Transactions," Dept. Comput. Sci., Univ. Minnesota, Tech. Rep. TR96-050, Mar. 1997

[6] C. Romero, S. Ventura "Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007

[7] R.Cooley, B. Mobasher, and J. Srivastava. "Web mining: Information and Pattern Discovery on the World Wide Web." *In Proceedings of Ninth IEEE International Conference on Tools with Artificial Intelligence" (ICTAI'97),* November 1997.

[8] Bourlas, P., Giakoumakis, E., and Papakonstantinou, G. (1999). A Knowledge Acquisition and management System for ECG Diagnosis. Machine Learning and Applications: Machine Learning in Medical Applications. Chania, Greece, pp. 27-29.

[9] Icks A et al. Incidence of lower-limb amputations in the diabetic compared to the non-diabetic population. Findings from nationwide insurance data, Germany, 2005-2007. Experimental and Clinical Endocrinology & Diabetes, 2009, 117:500–504.

[10] Han Jiawei, Micheline Kamber, Data Mining: Concepts and Technique. Morgan Kaufmann Publishers,2000

AUTHOR'S PROFILE

Md. Ezaz Ahmed

Pursuing Ph.D. under the supervision of Prof, (DR.) Yogesh K. Mathur. and co-supervision of DR. Varun kumar. Currently, he is working with itm University as Asst. Professor. He did his M.E (CSE) in first division with honors. He has more than 17 years of experience out of which 15 years teaching and 2 years industry experience. He has published 12 research papers, 2 in international Journal others in national conference and in departmental journal. His area of interest includes Web Development, Web Mining, Data Mining Software Engineering, Software verification validation and testing, and Basics of computer and C programming. He is a member of Indian Society of Technical Education (ISTE). Co-author of one project book published in 1998. He has 1 project book, 3 lab manuals to his credit.

Dr. Varun Kumar

Ph.D. (Computer Science), Head , CSE Deptt., MVN Engineering College, Palwal, Haryana ,India. Presently 3 Ph. D students are working under his supervision. Dr. Varun Kumar, completed his PhD in Computer Science. He received his M. Phil. in Computer Science and M. Tech. in Information Technology. He has 13 years of teaching experience. He is recipient of Gold Medal at his Master's degree. His area of interest includes Data Warehousing, Data Mining, and Object Oriented Languages like C++, JAVA, C# etc. He has published more than 35 research papers in Journals/Conferences/Seminars at international/national levels. He is working as an Editorial Board Member / Reviewer of various International Journals and Conferences. He has 3 books, 5 study materials and 3 lab manuals to his credit.

Dr. Y. K. Mathur

Ph.D. (Theoretical Physics) - Moscow University, Moscow, Russia (1982), M.Sc. Physics and Mathematics – Moscow University, Moscow, Russia (1978). Post Doctoral Positions held: Joint Institute for Nuclear Investigation, DUBNA, Russia (April 1982-Feb.1983), Department of Physics, University of Rochester, USA(Feb.1983-Dec.1983) and Department of Physics, University of Bielefeld, Germany (Dec.1983-Dec.1984). Academic Positions Held: CSIR Pool Officer, Department od Physics and Astrophysics, University of Delhi (as a member of Quark Physics Team headed by Prof. A.N.Mitra (FNA)) (Jan.1985-Jan. 1986), CSIR Research Scientist , Department of Physics and Astrophysics, University of Delhi.(Jan.1986-Aug.1988), Lecturer (Astt. Professor), Department of Physics and Astrophysics, University of Delhi (August 1988-May 1994), Reader (Associate Professor), Department of Physics and Astrophysics, University of Delhi(December1994-May 2005), Professor, Department of Applied Sciences and Humanities, ITM, Gurgaon Delhi(July 2005 –January 2011), Head ASH and School of Physics (April, 2010 to January 2011) and Professor, Department of Applied Sciences, PDM college of Engineering, Bahadurgarh (January 2011 till date).
Teaching Experience : Post Graduate: 21years, at the Department of Physics and Astrophysics, University of Delhi
Undergraduate : 6years (5 years and 6 months at ITM, Gurgaon and 6 months at PDM college of Engineering, Bahadurgarh) Haryana, India.