International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 22, 2017 ISSN: 2348-6600
Reference ID: IJCS-263 PAGE NO: 1678-1683
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Study on Data Mining Techniques for Diabetes Mellitus

R. Manimaran1
1Research Scholar, PG and Research Departmet of Computer Science,
J.J.College of Arts and Science (Autonomous)
Email: rmanimaran_jjc@rediffmail.com

Dr. M.Vanitha2
2Assistant professor, Department of Computer Applications,
Alagappa University, Karaikudi, India
Email: mvanitharavi@gmail.com

Abstract – Data mining is an analytic process designed to explore large amount of data in search of consistent patterns and systematic relationships between variables and then to validate the findings by applying the detected patterns to new sub sets of data. The ultimate goal of data mining is prediction and predictive data mining is the most common type of data mining. The process of data mining consists of three stages 1) Initial exploration 2) Model building and 3) Deployment i.e., the application of the model to new data in order to generate prediction. In this paper, we discovered the unseen but important risk factors of diabetes and their relationships. We used the above mentioned three stages of data mining to identify how different risk factors contribute towards diabetes and how diabetes contribute to other diseases.

Keywords: Attribute Reduction, Decision tree induction, Gini index.

I. INTRODUCTION

Diabetes Mellitus [1] is a chronic metabolic disease. The chronic hyperglycemia of diabetes is associated with long-term complications such as neuropathies, nephropathies and retinopathies. The vast majority of cases of diabetes fall into two broad categories Type 1 Diabetes Mellitus (T1DM) [1] and Type2 Diabetes Mellitus (T2DM) [2]. The cause of T1DM is an absolute deficiency of insulin secretion which leads to an uncontrolled increase of blood glucose levels. T2DM is a much more prevalent category. The cause is a combination of resistance to insulin action and an inadequate compensatory insulin secretor response. Recently, a number of computer-based expert systems for supporting decision making in diabetes management have been proposed. The physiological interactions regarding diabetes are multidimensional, nonlinear, stochastic, time-variant, and patient specific, artificial neural networks. According to the research, the factors of diabetes are: being overweight, carrying fat around

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 22, 2017 ISSN: 2348-6600
Reference ID: IJCS-263 PAGE NO: 1678-1683
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

waist and stomach, lack of exercise, no balanced diet, age more than 45 years, having a family history of diabetes, stress, having high blood pressure(140/90 mm/Hg or higher), having high cholesterol level and so on. The main triggering mechanism for diabetes is hereditary. Researchers record that there is a risk of 40%, if only one parent is a diabetic patient. If both parents are diabetic, then the risk is 99% and if one parent is diabetic and other from a diabetic family, then the risk is 70% [3].

Diabetes is a disease in which the body does not produce or properly use insulin. Insulin is a hormone that is needed to convert sugar, starches and other food into energy needed for daily life. There are two major types of diabetes: Type 1 - A disease in which the body does not produce any insulin, most often occurring in children and young adults. People with Type 1 diabetes must take daily insulin injections to stay alive. Type 1 diabetes accounts for 5 to 10 percent of diabetes. Type 2 - A metabolic disorder resulting from the body's inability to make enough, or properly use insulin. It is the most common form of the disease. Type 2 diabetes accounts for 90 to 95 percent of diabetes. Type 2 diabetes is nearing epidemic proportions, due to an increased number of older populations, and a greater prevalence of obesity and sedentary lifestyles. The main reason for diabetes is having a family history of diabetes. Lack of balanced diet leads to diabetics, being overweight, lack of exercise all leads to diabetics. Stress also plays a vital role for diabetes. Diabetes can lead to chest pain or heart diseases[4]. High blood pressure also increases the level of sugar in blood. The various symptoms can be swelling in feet, swelling in face, tiredness, excessive thirst, fatigue, frequent urination, sudden weight loss, blurred vision, itching, slow healing of wounds, numbness in feet, excessive hunger, giddiness, shoulder pain, irritability [5]. The presented approach will help the physicians or doctors with a series of services for the treatment support and assistance and it aims at mining the classification rules by decision tree induction methodology to the data warehouse [6].

The International Diabetes Federation has claimed that presently 246 million people are suffering from diabetes worldwide and this number is expected to increase up to 380 million by 2025 [7].

II. METHODOLOGY

The methodology involves: Preparation of questionnaires to collect information, applying preprocessing steps to prepare the database, apply classification rule to predict the class label. The initial step of the process is to prepare the questionnaire. The research is done on one thousand and twenty four patients. The various factors considered for the questionnaire are: gender, age, occupation, working hours, age of diabetes, family history, stress, sugar levels, Blood pressure, urea, cholesterol and various symptoms such as numbness, shoulder pain so on. The questionnaires are distributed to the patients and the data are collected. Each patient was asked to fill the questionnaire and based on the data collected, the data warehouse is prepared. The initial preprocessing step is done to remove the missing values and then the various algorithms are

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 22, 2017 ISSN: 2348-6600
Reference ID: IJCS-263 PAGE NO: 1678-1683
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

applied. Here we have nearly 25 attributes and in order to reduce the number of attributes we go for attribute reduction technique. For building the decision tree, first we find the attributes of higher rank and then build the tree.

The physicians are assisted with the help of classification rules derived from the decision trees. A decision tree [8] is a flowchart like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. ID3[9], the decision tree induction algorithm adopts an approach in which decision trees are constructed in a top-down manner.

Given a tuple X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules. During the tree construction, attribute selection measure are used to select the attribute that best partitions the tuples into distinct classes. When the trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning [8] attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

In the decision trees [10], the input data set has one attribute called class C that takes a value from K discrete values 1……k, and a set of numeric and categorical attributes A1……..Ap. The goal is to predict C given A1……..Ap. Decision tree algorithms automatically split numeric attributes Ai into two ranges and they split categorical attributes Aj into two subsets at each node. Splitting is generally based on the information gain ratio or the gini index. The final step involves pruning nodes to make the tree smaller and to avoid model overfit. In general decision trees have reasonable accuracy and are easy to interpret if the tree has a few nodes.

Decision trees classifiers [8] are so popular because it does not require any domain knowledge, it can handle high dimensional data, the learning and classification steps of decision trees are simple and fast. Decision trees classifiers have good accuracy. The various attributes considered for the questionnaire are: gender, age, occupation, working hours, age of diabetes, family history, stress, sugar levels, Blood pressure, urea, cholesterol and various symptoms such as numbness, shoulder pain so on. Here we have nearly 25 attributes and in order to reduce the number of attributes we go for attribute reduction technique. For building the decision tree, first we have to find the attributes of higher rank and then build the tree. Using the algorithm based on Gini index, we find the Gini index [9] as follows:

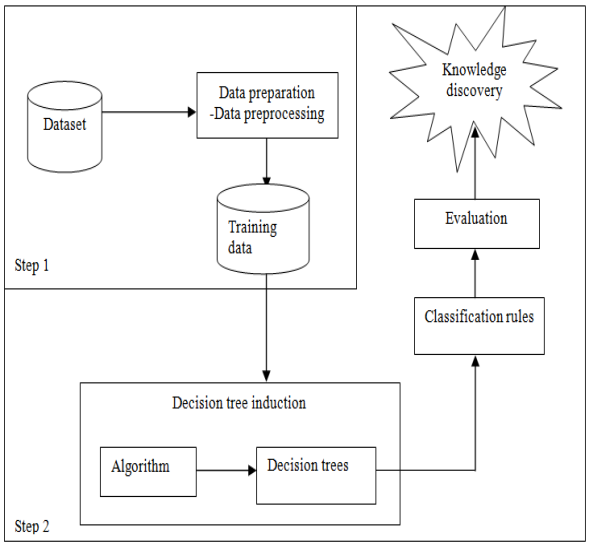
$$G(S)=1-\sum pi2$$

The various steps involved are: Calculate Gini index before split, Calculate Gini index after split, Gain= GI before split-GI after split. By setting a minimum threshold gain value, we select the highest ranked attributes. The attribute with highest gain is taken as the root node. Then we remove the highest gain attribute from the database and repeat the same procedure to build the decision tree. From

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 22, 2017 ISSN: 2348-6600
Reference ID: IJCS-263 PAGE NO: 1678-1683
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

the decision trees obtained, we can obtain the classification rules.

The dataflow is shown in Fig 1.



III. RESULTS AND DISCUSSION

Using the above mentioned methodology, we get the decision tress for sugar level before food and sugar level after food.

3.1 Sugar Level Before Food Among the 25 attributes, the relevant attributes are selected by Attribute Reduction technique [11],[12]. The attributes obtained by this technique are listed in Table 1. Here Cholesterol is having the highest gain, so it is chosen as root node and the procedure is repeated to construct the decision trees.

Table 1: Sugar Level Before Food-Relevant Attribute

Attribute	**Gain**
Age	0.1002
Age of diabetes	0.1258
Sugar after food	0.2005
Cholesterol	***0.2143***
Feet examination	0.1513
Nature of work	0.0887

3.2 Sugar Level After Food Among the 25 attributes, the relevant attributes are selected by Attribute Reduction technique. The attributes obtained by this technique are listed in Table 2.

Table 2: Sugar Level After Food-Relevant Attribute

Attribute	**Gain**
Age	0.1041
Sugar before food	0.1915
Cholesterol	***0.1820***
Feet examination	0.0944
Blood pressure	0.0793
Nature of work	0.0795

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 22, 2017 ISSN: 2348-6600
Reference ID: IJCS-263 PAGE NO: 1678-1683
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

IV.SUMMARY AND CONCLUSION

This study will help to create awareness about diabetic retinopathy among the people. We have discovered the risk factors of diabetes and the relationships. For sugar level before food, the risk factor is Cholesterol and sugar level after food, the risk factor is Sugar level before food. This study discovers that feet examination plays a very important role in examining the diabetic patients. Also stress is one of the reasons for diabetes. Diabetes may contribute to other diseases like heart attack, stroke.

In future, the work has to be extended as to predict knowledge using both classification and association concepts.

REFERENCES

[1] Eleni Georga, Vasilios Protopappas, 2009. "Data Mining for Blood Glucose Prediction and Knowledge Discovery in Diabetic Patients: The METABO Diabetes Modeling and Management System" 31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA, pp.5633-5636

[2] Patil.B.M, R.C.Joshi, Durga Toshniwal, 2010. " Association rule for classification of type-2 diabetic patients", in Proc.2nd IEEE International Conference on Machine Learning and Computing (ICMLC), pp. 330-334

[3] Hian Chye Koh and Gerald Tan, 2005. "Data Mining Applications in Healthcare", Journal of Healthcare information management, Vol. 19, No. 2, pp. 64-72.

[4] Reaven. G.M, 1988. "Role of insulin resistance in human disease", Diabetes, vol. 37, no.12, pp. 1595-1607.

[5] Pickup.J.C, G.Williams, (Eds.), Textbook of diabetes, Blackwell Science, Oxford.

[6] Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[7] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In InternetTechnology And Secured Transactions, 2012 International Conferece For, pp. 471-472. IEEE, 2012.

[8] Jiawei Han, Micheline Kamber, "Data mining: concepts and techniques",second edition.

[9] Ordonez C. 2006. "Comparing association rules and decision trees for disease prediction," in Proc.Int.Workshop Healthcare Inf.Knowl.Manag., pp.17–24

[10] Quentin.J, Truatvetter, P.Devos, A. Duharnel, R.Beuscan, 2002. "Assessing Association rules and decision trees on analysis of diabetes data from the DiabCare program in France" stud health technol inform, 90:557-61.

[11] Frank Lemke and Johann-Adolf Mueller, 2003. "Medical data analysis using self-organizing data mining technologies," Systems Analysis Modelling Simulation, Vol. 43, No. 10, pp: 1399 - 1408.

[12] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, 2005. "Knowledge Management, Data Mining, and Text Mining In Medical Informatics", Chapter 1, eds. Medical Informatics: Knowledge Management And Data Mining In Biomedicine, New York, Springer, pp. 3-34.

[13] Tarak Chaari, Frederique Laforest and Augusto Celentano, "Service-Oriented Context-Aware Application Design", LIRIS, INSA Lyon,

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 22, 2017 ISSN: 2348-6600
Reference ID: IJCS-263 PAGE NO: 1678-1683
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

Dipartimento di Informatica, Università Ca' Foscari di Venezia Via Torino 155, 30172 Mestre (VE), Italia.

[14] Theodor Chris Panagiotakopoulos, Dimitrios Panagiotis Lynras, Miltos Livaditis, Kyriakos N.Sgarbas, George C.Anastassopoulos and Dimitrios K.Lymberopoulos, 2010, "A Contextual

R. Manimaran1
1Research Scholar, PG and Research Departmet of Computer Science,
J.J.College of Arts and Science (Autonomous)
Email: rmanimaran_jjc@rediffmail.com

Dr. M.Vanitha2
2Assistant professor, Department of Computer Applications,
Alagappa University, Karaikudi, India
Email: mvanitharavi@gmail.com