



## DATA MINING TECHNIQUES FOR THE ANALYSIS OF TYPE-2 DIABETES—A SURVEY

V.SUDHARSANA

Research Scholar,

Department Of Computer Science,

PSG College of Arts & Science, Coimbatore, Tamilnadu, India.

[sudhamsc037@gmail.com](mailto:sudhamsc037@gmail.com)

T.REVATHI

Associate Professor,

[trevathi\\_psg@yahoo.co.in](mailto:trevathi_psg@yahoo.co.in)

**Abstract** - Invigour anxiety businesses, data mining plays an important role in early prophecy of diseases. Numerous tests must be conducted in a patient to detect a disease. Data mining techniques is used in disease prediction to reduce the test and increase the accuracy of rate of detection. Diabetes mellitus is one of the most common diseases among adolescent. This develops at a middle age and more common in overweight children and youngsters. so as to reduce the population with diabetes mellitus it should be detected at an earlier stage. This paper explore the survey on forecasting of diabetes using different data mining techniques such as Fuzzy Logic, Naive Bayes, J48 (C4.5), JRip, Neural networks, Decision trees.

**Keywords** - Diabetes Mellitus, Fuzzy Logic, Naive Bayes, J48 (C4.5), JRip, Neural networks, Decision tree.

### I. INTRODUCTION

Insulin is most essential hormone in the body. It changes over the sugar, starch and other nourishment things into vitality required for every

day life. In the event that the body does not create insulin the repetitive measure of sugar will be driven out by urination. This illness is Diabetes. Diabetes mellitus, commonly known as diabetes, is a group of metabolic diseases in which a person has high blood sugar. There are two reasons for high blood sugar: the pancreas does not produce enough insulin, or cells do not respond to the insulin that is produced. It is generally assessed that 347 million individuals all around have diabetes; and its frequency is expanding quickly. In this way there is an awesome need to concentrate on enhancing the forecast of diabetes. In this way, there is a need to build up a finding framework for diabetes with better accuracy. Data mining is the way toward breaking down information from alternate points of view and condensing it into helpful data that can be utilized for modern, medicinal and logical purposes. The present work has taken up to analyze the obtained data of diabetic patients by various data mining algorithms which can be helpful for medical analysts or practitioners for accurate diabetes diagnosis.

**Alagappa University, Karaikudi, India**

15<sup>th</sup> -16<sup>th</sup> February 2017

IT Skills Show & International Conference on Advancements in Computing Resources **(SSICACR-2017)**

<http://aisdau.in/ssicacr>

[ssicacr2017@gmail.com](mailto:ssicacr2017@gmail.com)

## II. DIABETES

The greater part of the food we eat is changed over to glucose, or sugar which is utilized for vitality. The pancreas secretes insulin which conveys glucose into the cells of our bodies, which thusly creates vitality for the ideal working of the body. When you have diabetes, your body either doesn't make a sufficient amount of insulin or can't utilize its own particular insulin and additionally it ought to. This causes sugar to develop in your blood prompting to intricacies like coronary illness, stroke, neuropathy, poor course prompting to loss of appendages, visual deficiency, kidney disappointment, nerve harm, and demise.

### Types of Diabetes

Type 1 - Diabetes likewise called as Insulin Dependent Diabetes Mellitus (IDDM), or Juvenile Onset Diabetes Mellitus is normally found in youngsters and youthful grown-ups be that as it may, more seasoned patients do give this type of diabetes every so often. In type1 diabetes, the pancreas experiences an immune system assault by the body itself in this manner; pancreas does not deliver the hormone insulin. The body does not appropriately metabolize food resulting in high (glucose) and the patient must depend insulin shots. Type I issue shows up in individuals more youthful than 35, typically from the ages 10 to 16.

Type II - Diabetes is likewise called as Non - Insulin Dependent Diabetes Mellitus (NIDDM), or Adult Onset Diabetes Mellitus. Patients deliver sufficient insulin however the body can't make utilization of it as there is an absence of affectability to insulin by the cells of the body.

Type II issue happens generally after the 40. Among the perpetual diabetic inconveniences, diabetic foot is the most obliterating outcome diabetes in India. Diabetes patients can regularly encounter loss of sensation in their feet. Indeed, even the littlest damage can bring about contamination that can be different genuine. 15% of patients with diabetes will create foot ulcers because of nerve harm and diminished blood stream. Diabetes gradually takes the individual's vision. It is the reason for basic visual deficiency and waterfalls. More than 50,000 leg removals happen each year because of diabetes in India. Diabetes patients can frequently encounter loss of sensation in their feet. Indeed, even the littlest harm can bring about contamination that can be different genuine. 15% of patients with diabetes will create foot ulcers because of nerve harm and lessened blood stream.

### General Symptoms of Diabetes

- Dehydration
- Frequent urination
- Weight loss
- Increased appetite
- Slow healing infections
- Blurred vision
- Unsettled stomach and vomiting

The following medical tests are used to diagnose the diabetic mellitus.

- Urine test
- Fasting blood sugar level
- Random blood glucose level
- Oral glucose tolerance test
- Glycosylated haemoglobin(HbA1c)



**Alagappa University, Karaikudi, India**

15<sup>th</sup> -16<sup>th</sup> February 2017

IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)

<http://aisdau.in/ssicacr>

[ssicacr2017@gmail.com](mailto:ssicacr2017@gmail.com)

### III RELATED STUDIES

So far a wide range of techniques have been advanced everywhere throughout the world for the forecast and treatment of diabetes. The most vital issue in all past canny techniques is the utilization of all traits for the conclusion of diabetes which other than being expensive reductions the level of exactness. It incorporates analyzing the publications, journals and reviews in the field of software engineering, designing, data mining and diabetes reports as of late. A five year test dataset is made to dig for information disclosure. Taking after data mining systems have been connected on diabetes information base.

Supervised machine learning algorithm:

The acquired information is arranged in view of different supervised machine learning algorithms, as Naive Bayes, Decision tree, ANN, JRip and J48(C4.5). TANAGRA a data mining tool for scholarly and research reason used to group the got information and assessed utilizing 10 - fold cross approval. It proposes a few data mining strategies from exploratory information investigation, measurable learning, machine learning and databases range.

Fuzzy logic:

The usage of computational insight procedure with fuzzy hierarchical model to show the strategy that restorative specialist use to perform early detection against DM in light of our survey. The model factors are classified into 3 types. Input variable, comprising of the estimation of polydipsia, polyphagia, polyuria, fasting blood glucose, 2-hour

post-prandial blood glucose and age from the user. At that point Temporary variable, comprising of the estimation of oral hazard that got from the fuzzy inference process amongst polydipsia and polyphagia. Atlast ,Output variable, is a variable that give data about the final result. The result of yield variable contains data about whether somebody has potential or no potential or high potential against DM.

Naive Bayes:

Naive Bayes classifier expect that the nearness or nonappearance of a specific component is irrelevant to the nearness or nonappearance of whatever other element, given the class variable. Applying Naive Bayes data mining classifier method which delivers an ideal forecast demonstrate utilizing least preparing set. Utilizing such characteristic, for example, age, sex, blood pressure and glucose and find the diabetic infection. Hindrance of Naive Bayes is emphasis are various, binning of constant contentions and high computational time.

J48 (C4.5):

J48 is an amplification of ID3. The extra elements of J48 representing missing qualities, decision trees pruning, nonstop property estimation ranges, induction of standards, and so forth. The goal is logically speculation of a choice tree until it picks up harmony of adaptability and exactness. At every node, pick the information that most successfully parts tests into subsets advanced in one class from the other. Then Set attribute with the most

**Alagappa University, Karaikudi, India**

15<sup>th</sup> -16<sup>th</sup> February 2017

IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)

<http://aisdau.in/ssicacr>

[ssicacr2017@gmail.com](mailto:ssicacr2017@gmail.com)

astounding standardized data pick up. Use this attribute to make a decision node and make the prediction.

Feature of C4.5 algorithm

- Handles preparing information with missing estimations of qualities
- Handles diverse cost qualities
- Pruning the decision tree after its creation
- Handles characteristics with discrete and consistent qualities

JRip:

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is one of the essential and most prevalent algorithms. Classes are analyzed in expanding size and introductory arrangement of standards for the class is produced utilizing incremental reduced error pruning. Assess RIPPER through JRip, an execution of RIPPER in WEKA with the parameters: folds=10; minNo=2; optimizations=2; seed=1; usepruning=true.

Decision tree:

It is a learning representation structure comprising of nodes and branches composed as tree to such an extent that, each inward non-leaf node is marked with estimations of the characteristics. Each node is marked with a class. Decision tree is a tree structure, which is type of stream outline. Utilizing nodes and internodes order and forecast are finished. Roots and internodes are utilized as experiments that differ the examples with various elements. Inner nodes are aftereffect of

trait cases. Leaf nodes signify the class variable. Class variable figure out whether individual has diabetes or not. Output of decision tree gives either tried positive or tried negative.

Neural Network:

An artificial neural network (ANN), frequently just called a "Neural network" (NN), is a numerical model or computational model in light of organic neural system. Neural network prepare data comparably the human mind does. Artificial neural network is best for pattern recognition. The dispersed information is sustained to information layer of ANN which comprise of n neurons. The hidden layer operation in view of input layer and it comprise of n+1. The yield layer represent either 0 and 1, here 0 utilized for patient not influenced by diabetes and 1 utilized for patient influenced by diabetes.

#### IV DATASET

The ability to diagnose diabetes early plays an important role for the patient's treatment process. The World Health Organization proposed the nine attributes, depicted below in Table I, of physiological measurements and medical test results for the diabetes diagnosis.

**TABLE-I: Attributes in Support Set**

S.No	Attribute
1.	Number of times pregnant

2.	Plasma glucose concentration in an oral glucose tolerance test
3.	Diastolic blood pressure(mm/Hg)
4.	Triceps skin fold thickness(mm)
5.	2-hour serum insulin( $\mu$ U/ml)
6.	Body mass index(kg/m)
7.	Diabetes Pedigree Function
8.	Age(year)
9.	Status(0-Healthy,1 Diabetes)

		Diabetic	Diabetic
1.	Diastolic B.P.(mm/Hg)	52-115	80-85
2.	Plasma glucose mg/dl	131-139	98-112
3.	Skin fold thikness	27-35	26-36
4.	BMI g/m <sup>2</sup>	27-32	26-30
5.	No of time pregnant	0-1	0-2
6.	2 hr serum InsulinU/ml	102-140	54-75

TABLE II: Comparison of various classifier

Classifier	Accuracy
Naive Bayes	95.85%
JRip	96.54%
J48 (C4.5)	100%
Decision tree	98.48%
Neural network	97.85%

TABLE III: Values of various attributes in supportset

S.No	Attribute	Support set	
			Non

TABLE IV: Analysis of classifier performance

Classifire	Time taken	Accuracy
Naive Bayes	845	55.85
JRip	765	65.48
J48 (C4.5)	658	68.58
Decision tree	875	52-58
Neural network	956	50-68

## V RESULTS AND DISCUSSION

For better comprehension comes about for every data mining techniques have appeared in different tables. Different classifiers are utilized as a part of mix with various data mining techniques for diabetes dataset analysis. Table I gives insights about different attributes chose for diabetes data analysis. It additionally indicates utilization of different information mining procedures to consider whether a patient can be analyzed high, low or medium for diabetes. Table II portrays the result of the exploration work by correlation finished with different classifiers. It was accounted for that J48 (C4.5) had defeated over different strategies by indicating 100% accuracy. J48 is extremely straightforward and precise classifier to settle on a choice tree over different classifiers. Table III demonstrates that the fuzzy and genetic algorithm produces fuzzy principles in view of support set. The outcomes got by supervised machine learning had demonstrated that the time taken for information investigation was high in ANN. The precision was high and time taken was slightest in J48 (C4.5). This demonstrates that the computational cost for information investigation was low in J48 (C4.5) thus the execution is exact (Table IV).

## VI. CONCLUSION

The revelation of information from medicinal datasets is critical with a specific end goal to make powerful therapeutic finding. The point of data mining is to focus learning from data put away in dataset and create clear and justifiable depiction of examples. Subsequently in conclusion it is

demonstrated that various data mining techniques were utilized to break down the got diabetes information. From this analyses shows that the cost for data analysis was stumpy in J48(C4.5) where the performance is better when compared to other data mining techniques.

## REFERENCES

- [1] Pardha Repalli, "Prediction on Diabetes Using Data mining Approach".
- [2] G. Parthiban, A. Rajesh, S.K. Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.
- [3] Lin, C., Lee, C., "Neural Fuzzy Systems," Prentice Hall, NJ, 1996.
- [4] Tsoukalas, L., Uhrig, R., "Fuzzy and Neural Approaches in Engineering," John Wiley & Sons, Inc., NY, 1997.
- [5] [http://en.wikipedia.org/wiki/Diabetes\\_mellitus](http://en.wikipedia.org/wiki/Diabetes_mellitus).
- [6] Cortes, C., Vapnik, V., "Support Vector Networks," Machine Learning, 20:273 - 297, (1995).
- [7] P. Radha, Dr. B. Srinivasan, "Predicting Diabetes by consequencing the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering &



**Alagappa University, Karaikudi, India**

15<sup>th</sup> -16<sup>th</sup> February 2017

IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)

<http://aisdau.in/ssicacr>

[ssicacr2017@gmail.com](mailto:ssicacr2017@gmail.com)

Technology, vol. 1 Issue 6, August 2014, pp. 334-339.

[8]Murat Koklu and YauzUnal, “ Analysis of a population of Diabeticpatients Databases with Classifiers”,International Journal of Medical,Health,Pharmaceutical and Biomedical Engineering”, vol.7 No.8, 2013.

[9]Riccardo Bellazzi and Ameen Abu-Hanna. 2009. Data mining technologies for blood glucose and diabetes management. Journal of diabetes science and technology. 3(3): 603-612. PMID: 20144300.

[10]ChaitraliDangare, S. and SulabaApte,S.”Improved study of disease prediction using data mining classification techniques”.Int.J.Comp.Appl.,2012,47(10):75-88.

[11]Yang Guo,GuohuaBai,Yan Hu School of computing Blekinge Institute of Technology Karlskrona,Sweden, “Using Bayes Network for prediction of Type-2 Diabetes”.