An Empirical Study about Type2 Diabetics using Duo mining Approach

[1]V.V.Jaya Rama Krishnaiah, [2]D.V. ChandraShekar,
[3] Dr. R. Satya Prasad, [4] Dr. K. Ramchand H Rao

[1]Associate Prof, Dept of CS, ASN College, Tenali,
[2]Associate Prof, Dept of CS, TJPS COLLEGE, GUNTUR,
[3]Associate Professor, Department of CSE, Acharya Nagarjuna University, Guntur,
[4]Professor, Department of CSE, ASN Womens Engineering College, Nelapadu,

Abstract:

Due to the revolutionary change in data mining and bio-informatics, it is very useful to use data mining techniques to evaluate and analyze bio-medical data. In this paper we propose a frame work called duo-mining tool for intelligent Text mining system for diabetic patients depending on their medical test reports. Diabetes is a chronic disease and major problem of morbidity and mortality in developing countries. The International Diabetes Federation estimates that 285 million people around the world have diabetes. This total is expected to rise to 438 million within 20 years. Type-2 diabetes mellitus (T2DM) is the most common type of diabetes and accounts for 90-95% of all diabetes. Detection of T2DM from various factors or symptoms became an issue which was not free from false presumptions accompanied by unpredictable effects. According to this context, data mining and machine learning could be used as an alternative way help us in knowledge discovery from data. We applied several learning methods, such as K-Nearest Neighbor, decision tree, support vector machines, acquire information from historical data of patient's from medical practicing centers in and around Guntur. Rules are extracted from Decision tree to offer decision-making support through early detection of T2DM for clinicians. Through this paper, we tried to determine how the extracted knowledge by the Text Mining is integrated with expert system knowledge to assist crucial decision making process.

Keywords: Text Mining, K-Nearest Neighbor, Support Vector Machines, Decision Trees, Type-2 diabetes, Duo Mining, Data Mining.

1. Introduction

Diabetes is an illness which occurs as a result of problems with the production and supply of insulin in the body [1]. People with diabetes have high level of glucose or "high blood sugar" called hyperglycemia. This leads to serious long-term complications such as eye disease, kidney disease, nerve disease, disease of the circulatory system, and amputation this is not the result of an accident.

Diabetes also imposes a large economic impact on the national healthcare system. Healthcare expenditures on diabetes will account for 11.6% of the total healthcare expenditure in the world in 2010. About 95% of the countries covered in this report will spend 5% or more, and about 80% of the countries will spend between 5% and 13% of their total healthcare dollars on diabetes [2]

Type-2 diabetes mellitus (T2DM) is the most common type of diabetes and accounts for 90-95% of all diabetes patients and most common in people older than 45 who are overweight. However, as a consequence of increased obesity among the young, it is becoming more common in children and young adults [1]. In T2DM, the pancreas may produce adequate amounts of insulin to metabolize glucose (sugar), but the body is unable to utilize it efficiently. Over time, insulin production decreases and blood glucose levels rise. T2DM patients do not require insulin treatment to remain alive, although up to 20% are treated with insulin to control blood glucose levels [3].

Diabates has no obvious clinical symptoms and not been easy to know, so that many diabetes patient unable to obtain the right diagnosis and the treatment. Therefore, it is important to take the early detection, prevent and treat diabetes disease, especially for T2DM.

Recent studies by the National Institute of Diabetes and Digestive and Kidney Diseases (DCCT) in India shown that effective control of blood sugar level is beneficial in preventing and delaying the progression of complications of diabetes [4]. Adequate treatment of diabetes is also important, as well as lifestyle factor such as smoking and maintaining healthy bodyweight [3].

According to this context, data mining and machine learning could be used as an alternative way in discovering knowledge from the patient medical records and classification task has shown remarkable success in the area of employing computer aided diagnostic systems (CAD) as a "second opinion" to improve diagnostic decisions [5]. In this area, classifier such as SVMs have demonstrated highly competitive performance in numerous real-world application such medical diagnosis, SVMs as one of the most popular, state-of-the-art data mining tools for data mining and learning [6].

In modern medicine, large amount of data are collected, but there is no comprehensive analysis to this data. Intelligent data analysis such as data mining was deployed in order to support the creation of knowledge to help clinicians in making decisions. The role of data mining is to extract interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from large amounts of data, in such a way that they can be put to use in areas such as decision support, prediction and estimation [7].

With this paper, we make two contributions. We present empirical result of inductive methods for detecting T2DM using machine learning and data mining. We structured the paper as follow: section 2 provides a brief explanation about several classifiers and medical data used in this research and the detailed information about text tokens were also explained in this section. Section 3 gives experimental design, and experimental result and discussion; we concluded the paper with summarization of the result by emphasizing this study and further research.

2. Research Method

2.1. Data Collection

We collected diabetic's patients from one of the government public hospital NRI and Manipal Institute , Guntur-Andhra Pradesh, India from 2009 to 2011. The patients included only type-2 diabetes, whereas other types of diabetes were excluded. All patients of this database are men and women of age between 0-20 Years. The variable takes the value "TRUE" and "FALSE", where "TRUE" means a positive test for T2DM and "FALSE" means a negative test for T2DM.
It is important to examine the data with preprocessing which consist of cleaning, transformation and integration. The analysis clinical attributes: (1) Gender, (2) Body mass, (3) Blood pressure, (4) Hyperlipidemia or Cholesterol (5) Fasting blood sugar (FBS), (6) Instant blood sugar, (7) Family history, (8) Diabetes Guest history, (9) Habitual Smoker, (10) Plasma insulin or Glucose, and (11) Age. The preprocessing method is briefly explained in next section.

2.2. Methodology

2.2.1 Text Mining

Text mining is also called as intelligent text analysis, text data mining, or knowledge discovery in text uncovers previously invisible patterns in existing resources. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining Text Mining itself is not a function, it combines different functionalities Searching, Information Extraction (IE), Categorization, Summarization, Prioritization, Clustering, Information Monitor and Information Retrieval. Information Retrieval (IR) systems identify the documents in a collection which match a user's query. The major steps in text mining were 1) Text Gathering and 2) Text Preprocessing. Text gathering includes collection of raw documents like Patient information which were in the text or script format and these documents may contain unstructured data. Preprocessing phase starts with tokenization. Tokenization is division of a document into terms. This process also referred as feature generation. In text preprocessing The raw data in the form of text files is collected from text scripts or Flat files. The data is converted in to structured format and stored in Microsoft Access Database.

To do the Information Extraction and Search, we use an mechanism call Duo-Mining developed in Java as the part of the Text Mining Tool and converts the Unstructured data into the structures manner.

The following Figure 1, shows the Text Mining and Conversion of unstructured data into structures data set.

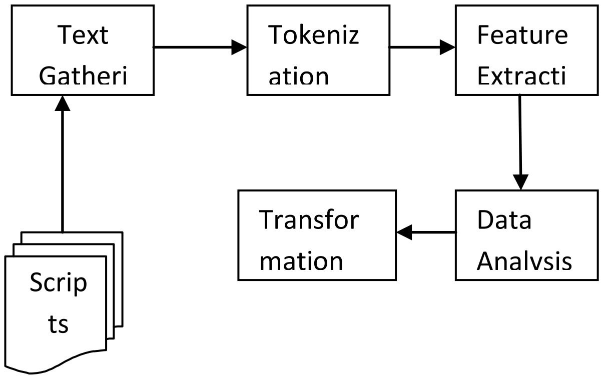
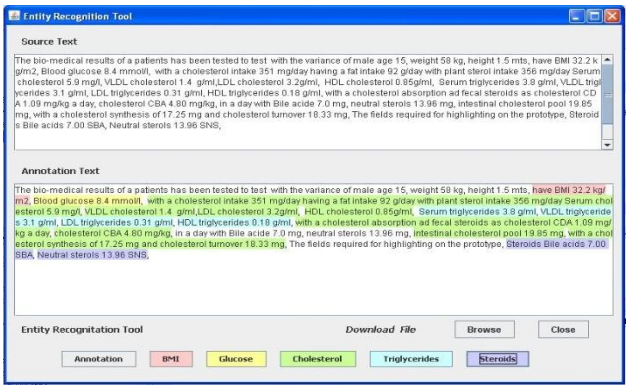


Figure 1: Conversion of Unstructured data into Structured Data Set using Duo-Mining Tool.



The bio-medical results of a patients has been tested to test with the variance of male age 15, weight 58 kg, height 1.5 mts, have BMI 32.2 kg/m2, Blood glucose 8.4 mmol/l, with a cholesterol intake 351 mg/day having a fat intake 92 g/day with plant sterol intake 356 mg/day Serum cholesterol 5.9 mg/l, VLDL cholesterol 1.4 g/ml,LDL cholesterol 3.2g/ml, HDL cholesterol 0.85g/ml, Serum triglycerides 3.8 g/ml, VLDL triglycerides 3.1 g/ml, LDL triglycerides 0.31 g/ml, HDL triglycerides 0.18 g/ml, with a cholesterol absorption ad fecal steroids as cholesterol CDA 1.09 mg/kg a day, cholesterol CBA 4.80 mg/kg, in a day with Bile acide 7.0 mg, neutral sterols 13.96 mg, intestinal cholesterol pool 19.85 mg, with a cholesterol synthesis of 17.25 mg and cholesterol turnover 18.33 mg, The fields required for highlighting on the prototype, Steroids Bile acids 7.00 SBA, Neutral sterols 13.96 SNS,

The bio-medical results of a patients has been tested to test with the variance of male age 15, weight 58 kg, height 1.5 mts, have BMI 32.2 kg/m2, Blood glucose 8.4 mmol/l, with a cholesterol intake 351 mg/day having a fat intake 92 g/day with plant sterol intake 356 mg/day Serum cholesterol 5.9 mg/l, VLDL cholesterol 1.4 g/ml,LDL cholesterol 3.2g/ml, HDL cholesterol 0.85g/ml, Serum triglycerides 3.8 g/ml, VLDL triglycerides 3.1 g/ml, LDL triglycerides 0.31 g/ml, HDL triglycerides 0.18 g/ml, with a cholesterol absorption ad fecal steroids as cholesterol CDA 1.09 mg/kg a day, cholesterol CBA 4.80 mg/kg, in a day with Bile acide 7.0 mg, neutral sterols 13.96 mg, intestinal cholesterol pool 19.85 mg, with a cholesterol synthesis of 17.25 mg and cholesterol turnover 18.33 mg, The fields required for highlighting on the prototype, Steroids Bile acids 7.00 SBA, Neutral sterols 13.96 SNS,

Figure 2: A Prototype Model of a Duo-Mining Tools

2.2.2. Support Vector Machines (SVMs)

Support vector machine (SVMs) are supervised learning methods that generate input- output mapping functions from a set of labeled training datasets. The mapping function can be either a classifiaction function or a regression function [6]. According to Vapnik [11], SVMs has strategy to find the best hyperplane on input space called the structural minimization principle from statistical learning theory.

Given the training datasets of the form $\{(x_1,c_1),(x_2,c_2),...,(x_n,c_n)\}$ where c_i is either 1 ("yes") or 0 ("no"), an SVM finds the optimal separating hyperplane with the largest margin. Equation (1) and (2) represents the separating hyperplanes in the case of separable datasets.

$w.x_i+b \geq +1$, for $c_i = +1$ ……………………………………(1)

$w.x_i+b \leq -1$, for $c_i = -1$ ………………………………..(2)

The problem is to minimize $|w|$ subject to constraint (1). This is called constrained quadratic programming (QP) optimization problem represented by:

minimize $(1/2)\ ||w||^2$

subject to $c_i(w.x_i - b) \geq 1$ ……………………………………(3)

Sequential minimal optimization (SMO) is one of efficient algorithm for training SVM [12]

2.2.3. K-Nearest Neighbors

One of the simplest learning methods is the Nearest Neibhor [13]. To classify an unknown instance, the performance element finds the example in the collection most similar to the unknown and returns the example's class label as its prediction for the unknown. Variants of this method, such as IBS$_k$, find the k most similar instances and return the majority vote of their class labels as the prediction.

2.2.4 Decision Tree

A decision tree is a tree with internal nodes corresponding to attributes and leaf nodes corresponding to class labels. Most implementations use the gain ratio for attribute selection, a measure based on the information gain [12].

The following Table 1 illustrates the data specifications

Table 1: Data Set for Diabetes Specification

NO	Attribute Name	Explanation
1	Plasma Insulin/Glucose	Glucose Concentration (high/Low)
2	FBS	Fasting blood sugar (mg/dl)
3	Body Mass Index (BMI)	Body mass of patients (Kg)
4	Blood Pressure	Blood Pressure in mmHg
5	IBS	Instant Blood Sugar (mg/dl)
6	AGE	Patient age (child, adult, old)
7	Diabetes Gestational	When pregnant women, who have never had diabetes before (Boolean)
8	Family History	Condition of abnormally elevated levels of any or all lipids and/or lipoproteins in the blood(Boolean)
9	Hyperlipidemia/Cholesterol	In the blood(Boolean)
10	Smoker	Patient's smoking habit (Boolean)
11	Gender	Patient's gender (male or female)

3. Analysis and Results

We conducted two experimental studies using our data collection described previously. We first applied all the classification methods the dataset collected, and we examined and validated the accuracy both in quantitative and qualitative measure. Quantitative measure is computed in percent, whereas qualitative measure is acceptance degree of patterns by clinicians.

Testing for type 2 diabetes in children
Criteria – To identify and group classes of people according to age, sex and race in the form of cluster patterns, based on various datasets, we also performed various types of comparison with health with respect to symptoms of diabetic's patients.
Overweight (BMI > 85th percentile for age and sex, weight for height > 85th percentile, or weight > 120% of ideal for height) Plus, Any two of the following risk factors:

1. Family history of type 2 diabetes in first- or second-degree relative
2. Race/ethnicity (Indian)
3. Signs of insulin resistance or conditions associated with insulin resistance (acanthosis, hypertension, dyslipidaemia, or PCOS)

The following Tables 2.1(a) to Table 7.1(b) illustrates the different comparisons between healthy people and Diabetes.

Table-2.1(a): (Diabetes Mellitus): Comparison of Anti Oxidants with GHb in healthy People between the age group of 0- 20 years

Patient code	Age	Sex	Blood Glucose (>120)	Calcium	Zinc	Copper	Folic acid	Homo-cysteine
1	4	F	216	8	45	75	4.45	17
2	9	F	158	7.8	60	69	3.98	19
3	11	M	182	8.1	57	62	4.01	14
4	14	M	172	9	52	70	4.06	27
5	16	F	201	8.4	59	74	3.76	16
6	18	M	177	8.3	44	68	5.29	13
7	19	F	208	7.8	52	76	3.18	39
8	20	M	182	7.5	49	79	4.97	15

Table-2.1(b) (Healthy): Comparison of Anti Oxidants with GHb in healthy People between the age group of 0- 20 years

Patient code	Age	Sex	Blood Glucose (<=120)	Calcium	Zinc	Copper	Folic acid	Homocysteine
1	5	M	98	9.3	80	105	5.98	6
2	8	F	102	9.2	85	110	6.2	8
3	10	F	79	9.5	86	106	6.51	8
4	15	M	110	9.8	90	101	6.16	9
5	20	M	120	9.6	92	112	6.05	7

Table-3.1(a) (Diabetis): Comparison of Thyroid function in Diabetes Mellitus Patients between the age group of 0- 20 years

Patient code	Age	Sex	T3	T4	TSH
1	4	F	0.98	5.62	1.09
2	9	F	0.85	6.24	2.64
3	11	M	0.82	6.24	2.5
4	14	M	1.01	7.25	3.18
5	16	F	0.48	3.52	15.9
6	18	M	1.09	6.85	2.51
7	19	F	0.71	8.1	1.42
8	20	M	1.12	5.97	1.08

Table-3.1(b) (Healthy): Comparison of Thyroid function in Healthy people between the age group of 0- 20 years

Patient code	Age	Sex	T3	T4	TSH
1	5	M	0.87	6.45	0.98
2	8	F	1.04	5.93	1.05
3	10	F	0.92	7.01	2.45
4	15	M	0.89	5.64	1.81
5	20	M	1.02	6.2	2.15

Table-4.1(a) (Diabetis): Comparison of Enzymes and GHb in Diabetes Mellitus Patients between the age group of 0- 20 years

Patient code	Age	Sex	SGPT	SGOT	Alk. PHO	Glycosilated haemoglobin (GHb)
1	4	F	17	29	418	8.2
2	9	F	28	31	214	6.6
3	11	M	24	11	129	7.3
4	14	M	34	36	98	7
5	16	F	21	28	254	7.8
6	18	M	19	25	208	7.1
7	19	F	34	24	205	8
8	20	M	32	34	365	7.3

Table-4.1(b) (Healthy): Comparison of Enzymes and GHb in Healthy Peoples between the age group of 0- 20 years

Patient code	Age	Sex	SGPT	SGOT	Alk. PHO	Glycosilated haemoglobin GHb)
1	5	M	22	26	148	4.9
2	8	F	25	30	152	5
3	10	F	31	34	160	4.4
4	15	M	27	25	132	5.3
5	20	M	20	24	141	5.6

Table-5.1(a) (Diabetis): Comparison of Lipids (Different type of cholesterols) in Diabetes Mellitus Patients between the age group of 0- 20 years

Patient code	Age	Sex	BMI	Cholesterol	HDL-Chol	Triglycrides	VLDL	LDL
1	4	F	Over weight	101	41	125	25	35
2	9	F	Over weight	132	38	148	30	64
3	11	M	Over weight	130	45	150	30	55
4	14	M	Over weight	148	40	129	26	82
5	16	F	Proportional	167	42	115	23	102
6	18	M	Over weight	180	41	127	25	114
7	19	F	Proportional	125	39	142	28	58
8	20	M	Over weight	184	39	150	30	115

Table-5.1(b) (Healthy): Comparison of Lipids (Different type of cholesterols) in Healthy People between the age group of 0- 20 years

Patient code	Age	Sex	BMI	Cholesterol	HDL-Chol	Triglycr Ides	VLDL	LDL
1	5	M	Thin	162	42	128	27	93
2	8	F	Proportional	158	40	110	22	96
3	10	F	Thin	170	38	132	26	106
4	15	M	Thin	165	40	108	22	103
5	20	M	Proportional	160	42	115	23	95

Table-6.1(a) (Diabetes): Comparison of Regular parameters in Diabetes Mellitus Patients between the age group of 0- 20 years

Patient code	Age	Sex	Macro Albunuria	Anti-Insulin	Hb % (in g/dl)	Micro albunuria
1	4	F	Normal	8.1	8.9	10
2	9	F	Normal	5.6	7.6	18
3	11	M	Normal	3.7	10.1	9
4	14	M	Traces	9.1	9.5	22
5	16	F	Normal	5.4	8.4	17
6	18	M	Normal	7.9	9.1	16
7	19	F	Normal	2.4	10.2	20
8	20	M	+	9.1	6.9	42

Table-6.1(b) (Healthy): Comparison of Regular parameters in Healthy people between the age group of 0- 20 years

Patient code	Age	Sex	Macro albunuria	Anti-Insulin	Hb % (in g/dl)	Micro albunuria
1	5	M	Normal	0.5	12.5	12
2	8	F	Normal	0.1	13.6	17
3	10	F	Normal	0.2	13.0	18
4	15	M	Normal	0.6	12.9	20
5	20	M	Normal	0.3	14.0	16

Table-7.1(a) (Diabetes): Comparison of Regular parameters in Diabetes Mellitus Patients between the age group of 0- 20 years

Patient code	Age	Sex	Blood Glucose	Urea	Creatinine	Albumin	Bilirubin
1	4	F	216	45	1.2	2	0.9
2	9	F	158	38	1.1	2.9	0.7
3	11	M	182	42	1	2.6	1.1
4	14	M	172	29	1.4	3.8	0.9
5	16	F	201	34	1.2	2.5	0.8
6	18	M	177	24	0.8	2.9	1
7	19	F	208	45	1.3	3.2	1
8	20	M	182	56	1.8	3.9	1.1

Table-7.1(b) (Healthy): Comparison of Regular parameters in Healthy people between the age group of 0- 20 years

Patient code	Age	Sex	Blood Glucose	Urea	Creatinine	Albumin	Bilirubin
1	5	M	98	18	0.7	3.6	1.8
2	8	F	102	24	0.6	3.8	1.2
3	10	F	79	30	1	3.2	1
4	15	M	110	22	0.9	3.5	0.8
5	20	M	121	28	1	3.8	1

The following Table 8 illustrates the Classification accuracy (%) of each spitted feature in the dataset

Table 8. Classification Accuracy (%)

Attribute	KNN	SVMs	DT	Average
Blood Sugar/ FBS	95,86	95,86	95,17	95,51
Plasmainsulin/ Glucose	95,40	96,55	96,78	95,86
Hyperlipidemia/Cholesterol	95,17	97,01	96,55	96,36
BMI	95,40	96,55	96,78	95,86
Average	95,45	95,49	96,32	

This research has four main outcomes regarding to detect T2DM. First, we start the assumption with Plasma insulin or Glucose and then BMI, Cholesterol and Blood Sugar. As compared the average classification accuracy of three mechanisms. Decision Tree mechanism shows the better performance. Surprisingly the average accuracy for Plasma insulin or Glucose and BMI were appearing same in all mechanisms.

Second, internist detected T2DM only by their experience. Thus, presumption attributes such as smoker and gestational history was avoided. Our study finds those attributes was found in many diabetic patients.
The following Table 9 illustrates about the extracted rules for determining the diabetics for the age group 0-20.

Table 9. Qualitative Measure in Detecting T2DM

PATTERN/RULES Extracted from Decision Tree Internist's acceptance (Yes/No)	
R1 : If Plasmainsulin is high and BMI is overweight and Hyperlipidemia is equal to 0 and family history is equal to 0 and smoker is equal to 1 then class YES	**YES**
R2 : Else IF plasmainsulin is high and BMI is proportional and hyperlipidemia is equal to 1 and IBS is greater than or equal to 200 mg/dl AND Class YES ELSE IF plasmainsulin is low AND FBS is less than or equal to 126 mg/dl then Class NO	**YES**
R3: AND blood pressure is greater than or equal to 140/90 mmHg AND IBS is less than or equal to 200 mg/dl THEN class NO ELSE IF plasmainsulin is low AND FBS is greater than or equal to 126 mg/dl AND BMI is proportional AND IBS is less than or equal to 200 mg/dl THEN class NO	**NO**
R4 : ELSE IF plasmainsulin is high AND BMI is thin AND FBS is greater than or equal to 126 mg/dl AND hyperlipidemia is equal to 1 THEN class Yes	**YES**
R5 : hyperlipidemia is equal to 1 AND IBS is greater than or equal to 200 mg/d AND FBS is greater than or equal to 126 mg/dl THEN class NO	**NO**

The following Figure 3 shows the decision tree for the above deduced Rules

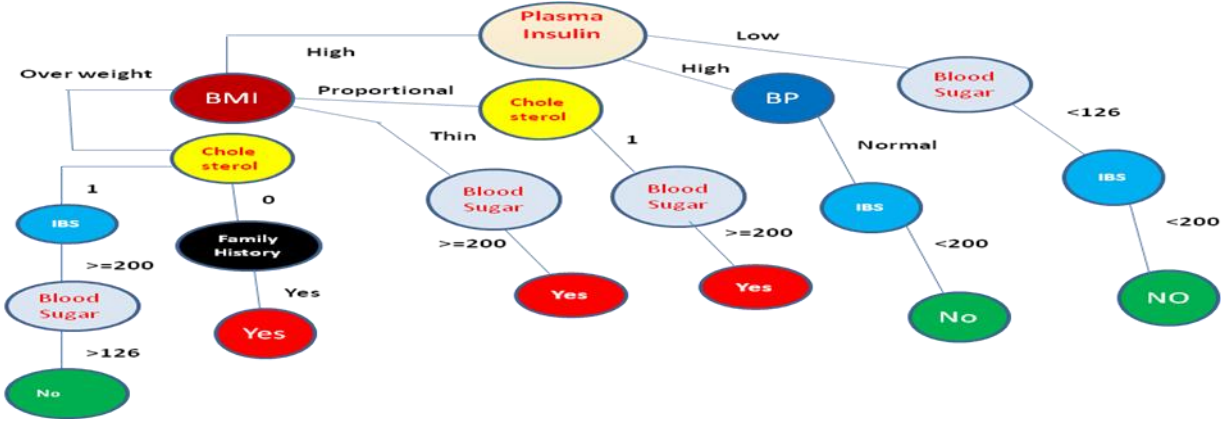


Figure 3: Decision Tree for Type 2 Diabetics for the age group 0-20

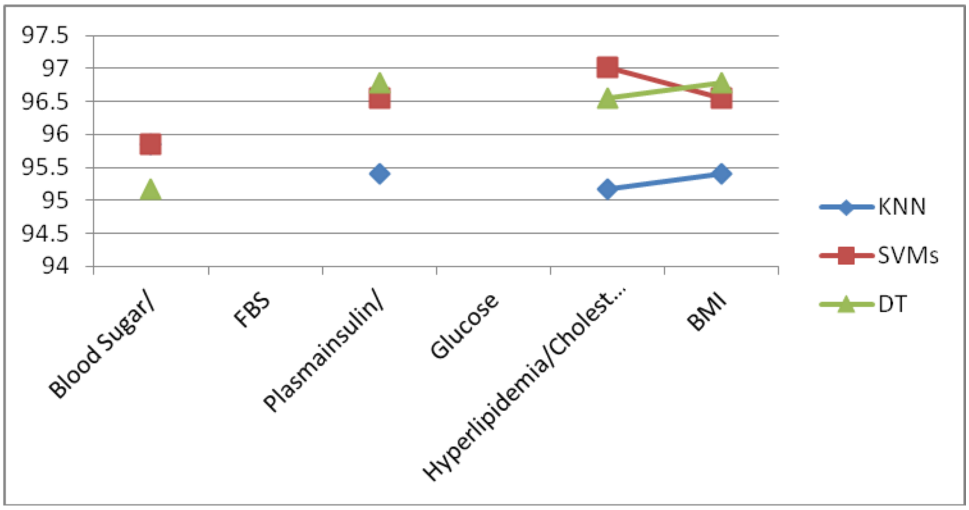
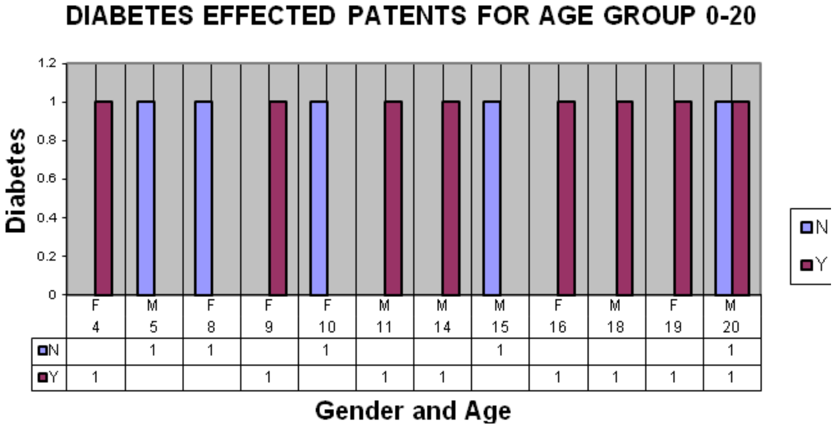


Figure 4: Accuracy Chart

The above graph gives a comparative report of predicting clinical trial of type 2 diabetics patients for age group 0-20 years, Decision tree method is the best prediction from the other two algorithms.

Our study involved every other children's hospital and NRI Hospital (n = 31), each diabetologist in private practice (n = 122), and every internal medicine unit (n = 164) in BW. A written clinical results and test with questionnaire and were used to identify children with T2DM and MODY who had been examined at any of these institutions between 2009 and 2011. Population data were drawn from the national census of 2011 and the subsequent annual updates.



	F 4	M 5	F 8	F 9	F 10	M 11	M 14	M 15	F 16	M 18	F 19	M 20
N		1	1		1			1				1
Y	1			1		1	1		1	1	1	1

Fig 5: Count of Diabetic Effected Persons for the Age Group 0-20 Years

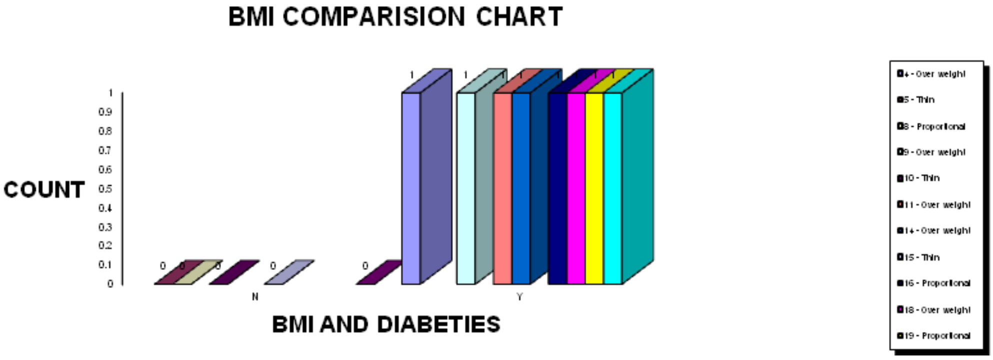


Fig 6 : BMI Comparison Cart for Age Groups 0-20

From the live chart (Figure 4, figure 5 and Figure 6), Type 2 diabetes is estimated to affect next over 4.5 million people in the Andhra Pradesh , India. The likelihood of developing type 2 diabetes is influenced by genetics and environment, If either parent has type 2 diabetes, the risk of inheritance of type 2 diabetes is 15%, If both parents have type 2 diabetes, the risk of inheritance is 75%, Almost 1 in 3 people with type 2 diabetes develops over kidney disease, Within 5 years of diagnosis of type 2 diabetes, 60% of people diagnosed have some degree of retinopathy, type 2 diabetes carries the risk of diabetes complications over a long period of time.

The major complication effect the patients with type 2 diabetes from the above three graphs :

Retinopathy. Up to 20%, most commonly occurs after the onset of puberty and after 5 to 10 years of diabetes duration, it has been reported in prepubertal children and with diabetes duration of only 1 to 2 years. Referrals should be made to eye care professionals with expertise in diabetic retinopathy, an understanding of the risk for retinopathy in the pediatric population, as well as experience in counseling the pediatric patient and family on the importance of early prevention/intervention. For children with type 2 diabetes, the first ophthalmologic examination should be obtained once the child is 10 years of age or older and has had diabetes for 3 to 5 years.

In type 2 diabetes, the initial examination should be shortly after diagnosis. In type 1 and type 2 diabetes, annual routine follow-up is generally recommended. Less frequent examinations may be acceptable on the advice of an eye care professional.

Nephropathy. – 12% of the patients were affected, To reduce the risk and/or slow the progression of nephropathy, optimize glucose and blood pressure control. In type 2 diabetes, annual screening should be considered at diagnosis. Screening may be done with a random spot urine sample analyzed for microalbumin-to-creatinine ratio. Confirmed, persistently elevated microalbumin levels should be treated with an ACE inhibitor, titrated to normalization of microalbumin excretion if possible.

Neuropathy. Although it is unclear whether foot examinations are important in children and adolescents, annual foot examinations are painless, inexpensive, and provide an opportunity for education about foot care. The risk for foot complications is increased in people who have had diabetes over 10 years.

Lipids. Based on data obtained from studies in adults, having diabetes is equivalent to having had a heart attack, making diabetes a key risk factor for future cardiovascular disease.

In children older than 2 years of age with a family history of total cholesterol over 240 mg/dl, or a cardiovascular event before age 55, or if family history is unknown, perform a lipid profile after diagnosis of diabetes and when glucose control has been established. If family history is not a concern, then perform a lipid profile at puberty.

Children with type 2 diabetes should have a lipid profile soon after diagnosis when blood glucose control has been achieved and annually thereafter. Experts also recommend lipid testing every two years if the lipid profile is normal.

To assess the prevalence of type 2 diabetes mellitus (T2DM) and Maturity onset diabetes of the young (MODY) in children and adolescents aged 0-20 yr in Guntur-Vijawayada(AP), India, and to compare our results with various algorithm techniques.

4. Conclusions:

The prevalence of T2DM for the age range from 0 to 20 yr is 2.30/100 000, whereas the prevalence of MODY in the same age range is 2.39/100 000. The median age of patients with T2DM was 15.8 yr, and 13.9 yr for MODY patients. The majority of patients with either T2DM or MODY were treated in children's hospitals and by consultant diabetologists. A molecular genetic analysis was done to substantiate the clinical diagnosis in less than half of the recruits (14.3% of T2DM and 44.8% of MODY patients). The prevalence of T2DM and MODY is considerably lower than the prevalence of type 1 diabetes. Type 2 diabetes thus continues to be a rare disease in children and adolescents in Andhra Pradesh, as is also the case in other states of India.

This paper collects and analyzes medical patient record of type-2 diabetes mellitus (T2DM) with knowledge discovery techniques to extract the information in the form of text or script format from T2DM patients in one of public hospital in Guntur, Andhra Pradesh, India. The experiment has successfully performed with several data mining techniques and Decision Tree mechanism as part of data mining technique achieves better performance than other classical methods such as K-NN Method. Extracted rules using decision tree are conformed to clinician's knowledge and more importantly, we found some major attributes such as plasmainsulin, BMI became a significant factor in our case study.This research might have some limitations and is being optimized. Later, it will focus on increasing the datasets in order to maximize result and discover novel optimal algorithm. As further researches, it would interest to include other risk factors such as sedentary lifestyle, Familiy History, Blood Pressure and Smoker.

References

[1] International Diabetes Federation (IDF), What is diabetes?, World Health Organisation, accessed January 2010, http://www.idf.org

[2] Zang Ping, et al. Economic Impact of Diabetes, International Diabetes Federation, accessed January 2010, http://www.diabetesatlas.org/sites/default/files/Economic% 20impact%20of%20Diabetes.pdf.

[3] Holt, Richard I. G., et al, editors. Textbook of Diabetes. 4th ed., West Sussex: Wiley-Blackwell; 2010.

[4] National Diabetes Information Clearinghouse (NDIC), The Diabetes Control and Complications Trial and Follow-up Study, accessed January 2010, http://diabetes.niddk.nih.gov/dm/pubs/control.

[5] N. Lavrac, E. Keravnou, and B. Zupan, Intelligent Data Analysis in Medicine, in Encyclopedia of Computer Science and Technology, vol.42, New York: Dekker, 2000.

[6] Olson, David L and Dursun Dulen. Advanced Data Mining Techniques, Berlin: Springer Verlag, 2008.

[7] Huang, Y., et al. Feature Selection and Classification Model Construction on Type 2 Diabetic Patients' Data. Journal of Artificial Intelligence in Medicine, 2007; 41: 251-262.

[8] Barakat, et al. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. IEEE Transactions on Information Technology in BioMedicine, 2009.

[9] Polat, Kemal and Salih Gunes. An Expert System Approach Based on Principal Component Analysis and Adaptive Neuro-Fuzzy Inference System to Diagnosis of Diabetes Disease. Expert System with Applications, Elsevier, 2007: 702-710.

[10] Yue, et al. An Intelligent Diagnosis to Type 2 Diabetes Based on QPSO Algorithm and WLSSVM. International Symposium on Intelligent Information Technology Application Workshops, IEEE Computer Society, 2008.
[11] Vapnik, V. The Nature of Statistical Learning Theory 2nd Edition, New York: Springer Verlag, 2000.
[12] Witten, I.H., Frank, E. Data mining: Practical Machine Learning Tools and Techniques 2nd Edition. San Fransisco: Morgan Kaufmann. 2005.
[13] Alpaydm, Ethem. Introduction to Machine Learning, Massachusetts: MIT Press, 2004: 154-155.
[14] Han, J. and Micheline Kamber. Data Mining: Concepts and Techniques, San Fransisco: Morgan Kaufmann Publisher, 2006: 310-311.
[15] Kohavi, R., Scaling Up the Accuracy of Naive Bayes Classifiers: A Decision Tree Hybrid, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996.
[16] Freund, Y., Schapire, R.E. Experiments with a New Boosting Algorithm. Proceedings of the Thirteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1996: 148–156.
[17] Opitz, D., Maclin, R.: Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research,*1999, 11: 169–198.
[18] Fawcett, Tom. An Introduction to ROC Analysis. *Pattern Recognition Letters*, Elsevier, 2006; 27: 861- 874.
[19] Zou, Kelly H. ROC literature research, On-line bibiography accessed February 2011, http://www.spl.harvard.edu/archive/spl-pre2007/pages/ppl/zou/roc.html
[20] V.V.Jaya Rama Krishnaiah, D.V.Chandra Sekhar, Dr. K.Ramchand H Rao, Dr. R Satya Prasad, Predicting the Diabetes using Duo Mining Approach, *International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 6, August 2012*

AUTHOR'S PROFILE

V.V.Jaya Rama Krishnaiah, received Master‘s degree in Computer Application from Acharya Nagrajuna University,Guntur, India, Master of Philosophy from Vinayaka University, Salem . He is currently working as Associate Professor, in the Department of Computer Science, A.S.N. Degree College, Tenali, which is affiliated to Acharya Nagarjuna University. He has 14 years teaching experience. He is currently pursuing Ph.D., at Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. His research area is Clustering in Databases. He has published several papers in National & International Journals.

D.V. Chandra Shekar, received Master of Engineering with Computer Science & Engineering He is currently working as Associate Professor, in the Department of Computer Science, T.J.P.S COLLEGE (P.G COURSES),Guntur, which is affiliated to Acharya Nagarjuna University. He has 14 years teaching experience and 1 years of Industry experience. He has published 52 papers in National & International Journals.

Dr. R.Satya Prasad, received Ph.D in Computer Sceicne in the faculty of Engineering in 2007 from Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. He have a satisfactory consistent academic track of record and received Gold medal from Acharya Nagarjuna University for his outstanding performance in a first rank in Masters Degree. He is currently working as Associate Professor in the Department of Computer Science and Engineering.

Dr.K Ramchand H Rao, received Doctorate in from Acharya Nagarjuna University, Master‘s degree in Technology with Computer Science from Dr. M.G.R University, Chennai, Tamilnadu, India. He is currently working as Professor and Head of the Department, Department of Computer Science and Engineering, A.S.N. Women's Engineering College, Tenali, which is affiliated to JNTU Kakinada. He has 18 years teaching experience and 2 years of Industry experience at Morgan Stanly, USA as Software Analyst. His research area is Software Engineering. He has published several papers in National & International Journals.