ID3 ALGORITHM IN DATA MINING APPLIED TO DIABETES DATABASE

Dr.R.Jamuna

Professor, Department of Computer Science S.R.College,
Bharathidasan university, Trichy.

Abstract

The past decade has seen a flurry of promising breakthroughs in data mining predictions. Many of these developments hold the potential to prevent dreadful diseases to improve the quality of life. The advances in medical science and technology have corresponded to the use of computer algorithms as an intermediary between the medical researchers and technocrats. Diabetes Mellitus is a major killer disease of mankind today. Data mining techniques can be used to highlight the significant factors causing such a disorder. Even though total cure is not possible for this pancreatic disorder the complications can be avoided by awareness using data mining algorithms. In this paper, eight major factors playing significant role in the Pima Indian population are analyzed. Real time data is taken from the dataset of National Institute of Diabetes and Digestive and Kidney Diseases. The data is subjected to an analysis by logistic regression method using spss 7.5 statistical software, to isolate the most significant factors. Then the significant factors are further applied to decision tree technique called the Iterative Dichotomiser-3 algorithm which leads to significant conclusions. Conglomeration of data mining techniques and medical research can lead to life saving conclusions useful for the physicians.

Keywords: BMI, Diabetes, decision tree, logistic regression, plasma.

ID3 ALGORITHM IN DATA MINING APPLIED TO DIABETES DATABASE

Introduction

Diabetes Mellitus is a major killer disease of mankind today. Data mining techniques can be used to highlight the significant factors causing such a disorder. Even though total cure is not possible for this pancreatic disorder the complications can be avoided by awareness about the factors playing major role in the cause of this disorder using data mining algorithms. In this paper, eight major factors, Prg (No. of times pregnant), Plasma (Plasma glucose concentration in Salvia), BP (Diastolic blood pressure), Thick (Forceps skin fold thickness), Insulin (Two hours serum insulin), Body (Body Mass Index; weight/height), Pedigree (Diabetes pedigree function), Age (in years), Response (1: Diabetic 0: Non-Diabetic), playing significant role in the Pima Indian population are analyzed. Real time data is taken from the large dataset of http://www.niddk.nih.gov/ which is the home page for the National Institute of Diabetes and Digestive and Kidney Diseases. First the data is sampled by eliminating any record which has a zero value for any field from the total real time data base. Next the data is subjected to an analysis by Logistic regression method by using spss 7.5 statistical software to show the most significant factors among the eight factors taken. Then the significant factors are applied to a Iterative Dichotomiser-3 algorithm which generates Decision Trees using Shannon Entropy for further investigations. Decision tree technique called the ID3 algorithm of data mining leads to significant conclusions about this diabetes disorder which poses to be the greatest threat to mankind in the coming era. Conglomeration of data mining techniques and medical data base research can lead to life saving conclusions for the physicians at critical times to save the mankind.

1. METHODS OF BUILDING DECISION TREES IN DATA MINING

In data mining, a decision tree is a predictive model; that is, a mapping of observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification tree or reduction tree. In these tree structures, leaves represent [1] classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or decision trees. In decision theory and decision analysis, a decision tree is a graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It can be used to create a plan to reach a goal. Decision trees are constructed in order to help with making decisions [2]. A decision tree is a special form of tree structure and a descriptive means for calculating conditional probabilities.

Decision tree learning is a common method used in data mining. Each interior node corresponds to a variable; an arc to a child represents a possible value of that variable. A leaf represents a possible value of target variable given the values of the variables represented by the path from the root. A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is either non-feasible, or a singular classification can be applied to each element of the derived subset. A random forest classifier uses a number of decision trees, in order to improve the classification rate. In data mining [3], trees can be described also as the synergy of mathematical and computing techniques that aids on the description, categorization and generalization of a given set of data. Data comes in records of the form:

$(X_i, y) = (x_1, x_2, x_3 ... x_k, y)$

The dependent variable, y, is the variable that we are trying to understand, classify or generalize. The other variables x_1, x_2, x_3 etc. are the variables that will help us for predictions.

1.1. Decision Trees in Data mining [10]

- An internal node is a test on an attribute.
- A branch represents an outcome of the test.
- A leaf node represents a class label.
- At each node, one attribute is chosen to split training examples into distinct classes.
- A new case is classified by following a matching path to a leaf node.

1.2. Types of building Decision Trees [4]

- Top-down tree construction
  - At start, all training examples are at the root.
  - Partition the examples recursively by choosing one attribute each time like age, Pdf etc…
- Bottom-up tree pruning
  - Remove sub trees or branches, in a bottom-up manner, to improve the estimated accuracy on new cases.

1.3. Choosing the Splitting Attribute

- At each node, available attributes are evaluated on the basis of separating the classes of the training examples. A Goodness function is used for this purpose.
- Typical goodness functions:
  - Information Gain (ID3)

- Information Gain Ratio
- Gini Index

Take all unused attributes and count their entropy concerning test samples.

- Choose attribute for which entropy is minimum.
- Make node containing that attribute.

2. PRINCIPLES USED IN THE ANALYSIS OF LARGE DATASETS

2.1. Information Entropy by Claude Shannon

In information theory, the Shannon entropy [7] or information entropy is a measure of the uncertainty associated with a random variable. It can be interpreted as the average shortest message length, in bits, that can be sent to communicate the true value of the random variable to a recipient. This represents a fundamental mathematical limit on the best possible lossless data compression of any communication: the shortest average number of bits that can be sent to communicate one message out of all the possibilities is the Shannon entropy. [8]

Information is measured as follows:-

- Given a probability distribution, the information required to predict an event is the distribution's *entropy.*
- Entropy gives the information required in bits.

Formula for computing the entropy:

$$\mathbf{ShannonEntropy}\left(p_1, p_2, ..., p_n\right) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n$$

2.2 Definition of Information Entropy

The information entropy of a discrete random variable X, that can take the range of possible values $\{x_1 \ldots x_n\}$ is defined to be,

$$H(X)=E(I(X))=\sum_{i=1}^{n} p(x_i)\log_2(1/p(x_i))$$

$$=\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

I(X) is the information content or self-information of X, which is itself a random variable; and $p(x_i) = P(X = x_i)$ is the probability mass function of X.

Introduced by Claude Shannon in 1948, ID3 (Iterative Dichotomiser-3) is an algorithm used to generate a decision tree. However, it does not always produce the smallest tree, and is therefore a heuristic. Occam's razor is formalized using the concept of information entropy as:

$$H(X)=-\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

2.2.1 Information Gain

- Uses Shannon Entropy
- IG calculates effective change in entropy after making a decision based on the value of an attribute.
- For decision trees, it's ideal to base decisions on the attribute that provides the largest change in entropy, the attribute with the highest gain.

2.3. Introduction to ID3 Algorithm for Diabetes Database

ID3 begins by choosing a random subset of the training instances. This subset is called the window. The procedure builds a decision tree that correctly classifies all instances in the window. The tree is then tested on the training instances outside the window. If all the instances are classified correctly then the procedure halts. Otherwise it adds some of the instances incorrectly classified to the window and repeats the process. This iterative

strategy is empirically more efficient than considering all instances at once. In building a decision tree ID3 selects the feature which minimizes the entropy function given below and thus best discriminates among the training instances. Data have been collected from about 768 Indian Origin females who were tested for the presence of diabetes mellitus of which 268 were found to be positive. Sample of 336 records are selected deleting the record sets with zero values.

3. RESULTS OF SOFTWARE BASED ANALYSIS OF DATASET.

3.1. Logistic Regression Outputs from SPSS7.5

Logistic regression method was applied to bring out the significance factors like age, obesity, etc. in the cause of the diabetes disorder, in Pima Indian diabetes database using SPSS 7.5 software. These factors are fuzzified to form a sample decision tree by ID3 algorithm.

Table 1: Logistic Regression Outputs from SPSS 7.5 [11]

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp (B)
AGE	.0408	.0192	4.5086	1	.0337	.0767	1.0416
BMI	.0761	.0315	5.7548	1	.0164	.0938	1.0791
BP	.0060	.0132	.2063	1	.6497	.0000	1.0060
INSU	.23E-05	.0014	.0005	1	.9822	.0000	1.0000
PDF	1.0970	.4776	5.2746	1	.0216	.0876	2.9951
PLAS	.0362	.0062	33.4697	1	.0000	.2717	1.0368
PRG	.0736	.0597	1.5197	1	.2177	.0000	1.0764
THICK	.0111	.0187	.3525	1	.5527	.0000	1.0112
Constant	-10.8272	1.4227	57.9161	1	.0000		

1. Results of Logistic Regression SPSS 7.5 version.

Total number of cases : 336
(Un-weighted)
Number of selected cases : 336
Number of unselected cases : 0
Dependent Variable Encoding:

Original value	Internal value
0.00	0
1.00	1

Dependent Variable : Response
Beginning Block Number 0.
Initial Log Likelihood Function -2 Log Likelihood 426.33781
* Constant is included in the model.
Beginning Block Number 1.
Method: Enter variable(s) Entered on Step Number

1.	Age	BMI	Bp
	Insulin	PDF	Plasma
	Prg	Thick	

Estimation terminated at iteration number 4 because Log Likelihood decreased by less than .01 percent.

2.	Log Likelihood	288.920
	Goodness of Fit	351.718
	Cox & Snell - R^2	.336
	Nagelkerke - R^2	.467

	Chi-Square	df	Significance
Model	137.417	8	.0000
Block	137.417	8	.0000
Step	137.417	8	.0000

Classification Table for Response
The Cut Value is .50

		Predicted .00	Predicted 1.00	Percent Correct
Observed	0	200	25	88.89%
.00	1	45	66	59.46%
		Overall		79.17%

From the observations of Table [1.1], we find that the following factors playing significant role in the cause of diabetes.

3.2. Analysis of Logistic Regression Results

- Age: sig = .0337 so 97% confidence level.
- Body Mass Index: sig = .0164 = .02 so 98% confidence level.
- PDF (Diabetes Pedigree Function): sig=.0216 so 98% confidence level. Implication of Hereditary Nature in the disease.
- Plasma (Glucose Concentration in Saliva): sig = .0000 100% confidence level as shown in Fig[1.1]

Table 2: Sample Dataset from Pima Indian diabetes database [5]

PATIENT	AGE	BMI	PLASMA	PDF	DIABETIC/ NOT
P1	YOUNG	HIGH	MEDIUM	LOW	NO
P2	YOUNG	HIGH	LOW	MEDIUM	NO
P3	YOUNG	NORMAL	MEDIUM	HIGH	YES
P4	MIDDLE	HIGH	HIGH	HIGH	YES
P5	OLD	HIGH	MEDIUM	HIGH	YES
P6	MIDDLE	HIGH	LOW	HIGH	NO
P7	OLD	NORMAL	HIGH	HIGH	NO
P8	OLD	HIGH	HIGH	HIGH	NO
P9	MIDDLE	NORMAL	LOW	HIGH	YES
P10	OLD	HIGH	MEDIUM	HIGH	YES
P11	YOUNG	HIGH	LOW	LOW	NO
P12	YOUNG	HIGH	MEDIUM	MEDIUM	NO
P13	MIDDLE	HIGH	HIGH	HIGH	NO
P14	YOUNG	NORMAL	MEDIUM	MEDIUM	NO
P15	MIDDLE	HIGH	HIGH	HIGH	NO

Table 3: Sample Fuzzified Range of Dataset. [5]

FUZZIFICATIONOF OF DATA SET-RANGE			
Age youth:21-38 middle:39 -81 large:>81	BMI: small:0-33 medium:34-52 large:>52	Plasma Glucose: small:128-158 medium:159-197 large:>197	Pdf: small:0.12-0.54 medium:0.54-2.33 large:>2.33

3.3 The ID3 Algorithm Applied to Diabetes Database [6]

1. Select a random subset W (called the "window") from the training set. Build a decision tree for the current window. Select the best feature which minimizes the entropy function H:

   $H = \sum -p_i \log p_i$(optimal values are available and the optimum entropy may be found by discrete probabilistic methods)

   Where p_i is the probability associated with i^{th} class. The entropy is calculated for each value. The sum of the entropy is calculated for each value. The sum of the entropy weighted by the probability of each value is the entropy for the feature. Categorize training instances into subsets by this feature. Repeat this process recursively until each subset contains instances of one kind (class) or some statistical criterion is satisfied.

2. Scan the entire training set for exceptions to the decision tree.
3. If exceptions are found, insert some of them into W and repeat from Step 2. The insertion may be done either by replacing some of the existing instances in the window or by augmenting it with the new exceptions. In practice a statistical criterion can be applied to stop the tree from growing as long as most of the instances are classified correctly. Fig [1.2]

3.3.1 ID3 ALGORITHM [9]

- Establish Classification Attribute as in Table [1.2].
- Compute Classification Entropy.
- For each attribute in R, calculate Information Gain using classification attribute.
- Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).

- Remove Node Attribute, creating reduced table R_S.
- Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced.

3.3.2 Establishing a Target Classification

Is the test patient diabetic?

- 5/15 yes, 10/15 no
- Calculating for the Classification Entropy

I_E= -(5/15)$\log_2$(5/15)-(10/15)$\log_2$(10/15) = ~0.918

3.3.3 Example – Information Gain for Age

- Age: 6 Young, 5 Middle, 4 Old
- 3 values for the attribute age, so we need 3 entropy calculations.

Table 4. Information Gain for Age

Information Gain for Age	Calculations
Young : 5 no, 1 yes	I_{young} = -(5/6) $\log_2$(5/6)-(1/6)$\log_2$(1/6) = ~0.65
Middle : 3 no, 2 yes	I_{middle} = -(3/5) $\log_2$(3/5)-(2/5) $\log_2$(2/5) = ~0.97
Old : 2 no, 2 yes	I_{old} = 1 (evenly distributed subset)

$$IG_{Age} = IE(S) - \left[(6/15)*I_{young} + (5/15)*I_{middle} + (4/15)*I_{old}\right]$$
$$IG_{Age} = 0.918 - 0.85 = 0.068$$

We must calculate Information Gain of remaining attributes to determine the root node.

3.3.4 Example - Information Gain for BMI

- 4 yes, 11 no
- 2 values for the attribute BMI, so we need 2 entropy calculations.

Table 5. Information Gain for BMI

Information Gain for BMI	CALCULATIONS
Yes : 2 yes, 2 no	I_{BMI} = 1 (evenly distributed subset)
No : 3 yes, 8 no	I_{LOWBMI} = -(3/11)log$_2$(3/11)-(8/11)log$_2$(8/11) = ~0.84

$$IG_{BMI} = IE(S) - \left[(4/15^*)I_{BMI} + (11/15)*I_{LowBMI}\right]$$
$$IG_{BMI} = 0.918 - 0.886 = 0.032$$

3.3.4 Example – Information Gain for Plasma

- 6 Middle, 4 Low, 5 High
- 3 values for attribute Plasma, so we need 3 entropy calculations

Table 6. Information Gain for Plasma

Information Gain for Plasma	CALCULATIONS
Medium : 3 no, 3 yes	I_{MIDDLE} = 1 (evenly distributed subset)
Low : 3 no, 1 yes	I_{LOW} = -(3/4)log$_2$(3/4)-(1/4)log$_2$(1/4) = ~0.81
High : 4 no, 1 yes	I_{HIGH}= -(4/5)log$_2$(4/5)-(1/5)log$_2$(1/5) = ~0.72

$$IG_{Plasma} = IE(S) - \left[(6/15)*I_{Medium} + (4/15)*I_{Low} + (5/15)*I_{High}\right]$$
$$IG_{Plasma} = 0.918 - 0.856 = 0.062$$

3.3.5 Example – Information Gain for PDF

- PDF: 2 Low, 3 Medium, 10 High
- 3 values for attribute PDF, so we need 3 entropy calculations.

Table 7. Information Gain for PDF

Information Gain for PDF	CALCULATIONS
Low : 0 yes, 2 no	$I_{low} = 0$ (no variability)
Medium : 0 yes, 3 no	$I_{medium} = 0$ (no variability)
High : 5 yes, 5 no	$I_{high} = 1$ (evenly distributed subset)

We can omit calculations for Low and Medium since thay always end up with not-diabetic category.

$$IG_{PDF} = IE(S) - \left[(10/15) * I_{high}\right]$$
$$IG_{PDF} = 0.918 - 0.667 = 0.248$$

3.3.6 Choosing the Root Node

Table 8. Finding Maximum Gain Factor

Finding Maximum Gain Factor	values
IG_{age}	0.068
IG_{BMI}	0.032
IG_{Plasma}	0.062
IG_{PDF}	0.248

Our best pick is PDF, and we can immediately predict the patient is not diabetic when PDF is Low or Medium. It also indicates that diabetes is a hereditary disease since PDF indicates the diabetes pedigree function from genes.

Fig-1 ROOT OF DECISION TREE

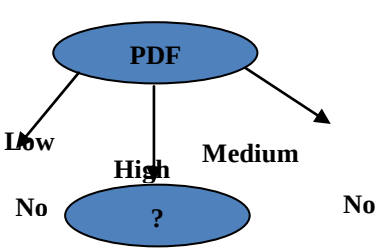


Table 1.3: Example – After Root Node Creation

Since we selected the PDF attribute for our Root Node, it is removed from the table for future calculations.

Table 9.Table of sample dataset after first iteration

PATIENT	AGE	BMI	PLASMA	PDF	DIABETIC/ NOT
P3	YOUNG	NORMAL	MEDIUM	HIGH	YES
P4	MIDDLE	HIGH	HIGH	HIGH	YES
P5	OLD	HIGH	MEDIUM	HIGH	YES
P6	MIDDLE	HIGH	LOW	HIGH	NO
P7	OLD	NORMAL	HIGH	HIGH	NO
P8	OLD	HIGH	HIGH	HIGH	NO
P9	MIDDLE	NORMAL	LOW	HIGH	YES
P10	OLD	HIGH	MEDIUM	HIGH	YES
P13	MIDDLE	HIGH	HIGH	HIGH	NO
P15	MIDDLE	HIGH	HIGH	HIGH	NO

4. ITERATION-II

Calculating for Entropy IE (PDF) we get 1, since we have 5 yes and 5 no.

4.1 Example – Information Gain for AGE

- Age: 1 Young, 5 Middle, 4 Old
- 3 values for attribute age, so we need 3 entropy calculations.

Table 10. – Information Gain for AGE

Information Gain for AGE	Calculations
Young : 1 yes, 0 no	$I_{young} = 0$ (no variability)
Middle : 2 yes, 3 no	$I_{middle} = -(2/5)\log_2(2/5)-(3/5)\log_2(3/5) = \sim 0.97$
Old : 2 yes, 2 no	$I_{old} = 1$ (evenly distributed subset)

$$IG_{age} = IE(S_{pdf}) - [(5/10) * I_{high} + (4/10) * I_{low}]$$
$$IG_{age} = 1 - 0.885 = 0.115$$

4.2 Example – Information Gain for BMI

- BMI: 3 yes, 7 no
- 2 values for attribute BMI, so we need 2 entropy calculations.

Table 11. Information Gain for BMI

Information Gain For BMI	Calculations
Yes : 2 yes, 1 no	$I_{BMI} = -(2/3)\log_2(2/3)-(1/3)\log_2(1/3) = \sim 0.84$
No : 3 yes, 4 no	$I_{LOWBMI} = -(3/7)\log_2(3/7)-(4/7)\log_2(4/7) = \sim 0.84$

$$IG_{BMI} = IE(S_{PDF}) - [(3/10) * I_{BMI} + (7/10) * I_{LowBMI}]$$
$$IG_{BMI} = 1 - 0.965 = 0.035$$

4.3 Example - Information Gain for Plasma

- Plasma: 3 Medium, 5 High, 2 Low
- 3 values for attribute weight, so we need 3 entropy calculations.

Table 12. Information Gain for Plasma

Information Gain For Plasma	Calculations
Medium : 3 yes, 0 no	$I_{Medium} = 0$ (no variability)
High : 1 yes, 4 no	$I_{High} = -(1/5)\log_2(1/5)-(4/5)\log_2(4/5) = \sim 0.72$
Low : 1 yes, 1 no	$I_{Low} = 1$ (evenly distributed subset)

$$IG_{Age} = IE(S_{PDF}) - \left[(5/10) * I_{High} + (2/10) * I_{Low}\right]$$
$$IG_{Age} = 1 - 0.561 = 0.439$$

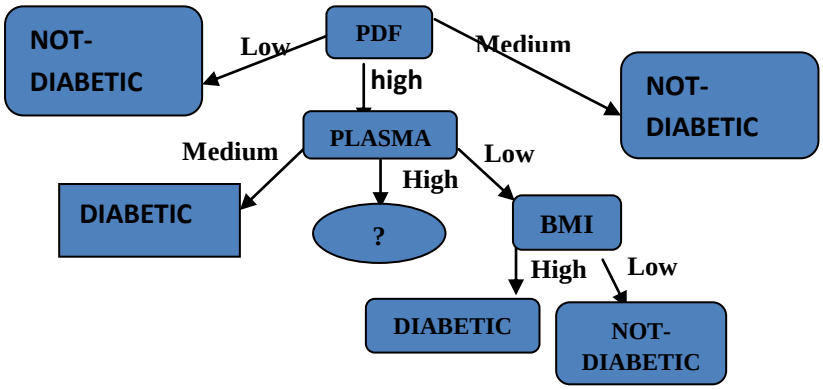
EXAMPLE - CHOOSING THE LEVEL 2 NODE

Table 13. Finding Maximum Gain Factor

Finding Maximum Gain Factor	Calculations
IG_{Age}	0.115
IG_{BMI}	0.035
IG_{Plasma}	0.439

- **Plasma has the highest gain, and is thus the best choice.**

Fig. 2: Example – Decision Tree: 1



Since there are only two items for BMI where Plasma =low and Plasma=medium the result is inconsistent.

Table 14: Example – Updated Table

AGE	BMI	DIABETIC/NOT
MIDDLE	HIGH	YES
OLD	NORMAL	NO
OLD	HIGH	NO
MIDDLE	HIGH	NO
MIDDLE	HIGH	NO

5. RESULTS AND DISCUSSIONS

- All patients with large BMI in the Fig [2] are diabetic.
- All patients with large age factor (OLD) in the Table [14] are not diabetic even with large BMI.
- Obesity or age alone cannot indicate the sure possibility of getting the disorder since hereditary factors and phenotypic factors like lifestyle, stress and habits play a role.
- Due to inconsistent patterns in the data, there is no way to proceed since middle age patients may be diabetic or non diabetic depending on other factors, so iterations are terminated.
- ID3 attempts to make the shortest decision tree out of a set of learning data, shortest is not always the best classification.
- Requires learning data to have completely consistent patterns with no uncertainty.

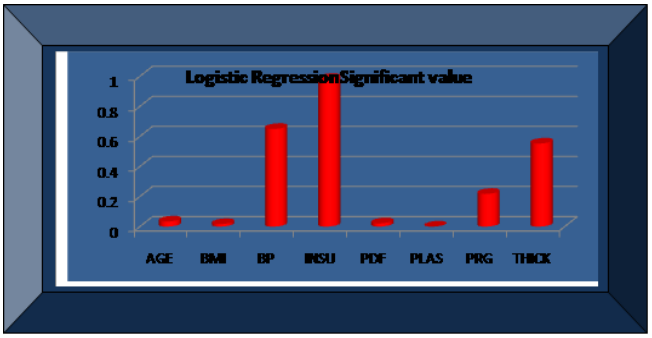
6. CONCLUSIONS

Data mining method using logistic regression implies that Age, Obesity, PDF and Plasma level are to be taken care of for the onset of diabetes mellitus. Pdf factor is the diabetes pedigree function value which shows the hereditary nature of the disorder which signifies that candidates with diabetes disorder in ancestors should take necessary monitoring of glucose levels periodically. ID3 algorithm applied to the sample database gives the decision tree prediction with major factors influencing diabetes. Pdf at first level of decision and Plasma glucose concentration in saliva at the next level and, the body mass index at the third level of decision have minimum entropy. Their significant role in the cause of diabetes is implied by the decision tree. Similarly larger decision tree can be drawn to show the significance of all factors which are responsible for the cause of diabetes by mining the total

dataset. The paper on a small scale tries to bring out the dominant factors alone by applying Iterative Dichotomiser ID3 algorithm of data mining. As our mankind has a great threat of this pancreatic disorder more in the coming era the sample data is chosen from diabetes database. The same idea can be applied to any disease database on a large sampling to bring out more useful diagnostic findings before complications affect the human population.

TABLES AND FIGURES

Figure 3: Significant Factors from Logistic Regression output



References

1. Ankerst, M., Elsen, C., Ester, M. and Kriegel, H.P. Visual classification: An interactive approach to decision tree construction. In Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99), San Diego, CA, Aug. 1999, pp. 392-396.
2. Almuallim H., an Efficient Algorithm for Optimal Pruning of Decision Trees. Artificial Intelligence, 1996, 83(2): 347-362.
3. Brodley C.E. and Utgoff. P.E., Multivariate decision trees, Machine Learning, 1995, 19: 45-77.
4. Quinlan, J. R (1985). Induction of Decision Trees, Machine Learning 1: 81-106, 1986.
5. http://www.niddk.nih.gov/ Home page for the National Institute of Diabetes and Digestive and Kidney Diseases.
6. Navathe Elmasri, (2007). Fundamentals of Database Systems (5th Edition), 975-977.
7. Shannon, Claude E. Prediction and Entropy of Printed English. (Retrieved 04/23/2010). http://languagelog.ldc.upenn.edu/myl/Shannon1.

8. Shannon, C.E. (1948). A mathematical theory of Shannon communication, Bell System Technical Journal **27**: 379-423 and 623-656. http://cm.bell labs.com/ cm/ms/what/ Shannon day/paper.html.
9. Ross, Peter (10/30/2000). Rule Induction: Ross Quinlan's ID3 Algorithm (Retrieved 04/23/2010). http://www.dcs.napier.ac.uk/~peter/vldb/dm/node11.html.
10. Quinlan, J.R., Simplifying decision trees, International Journal of Man machine Studies, 1987, num. 27, pp. 221-234.
11. R.Jamuna, K.Meena. Data mining by Logistic Regression Techniques in Pima Indian Diabetes Database. Bio-Science Research Bulletin, Vol. 22 (No. 2) July- December 2006.