Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

# The Role of Diabetes Mellitus in Prediction of Myocardial Infarction using CHAID based Data Mining Model

**Dr. S. Bharathidason[1] and C. Sujdha[2]**
[1]Assistant Professor and Head, Department of Computer Science, Loyola College (Autonomous), Chennai - 600 034.
[2]Assistant Professor, Department of Computer Science, Mar Gregorios College of Arts & Science, Chennai - 600 037.
**Corresponding authour:** bharathidasan@loyolacollege.edu

**Abstract**

The Classifiers are the mechanism to predict the possibilities of existence of the given data called test data with the collection of known values called trained data through machine learning. This paper is an attempt to identify the causative factor of diabetic's mellitus influencing Myocardial Infarction using most promising Classifier CHAID (Chi Square Automatic Interaction Detection) and also find its classification efficiency. The diabetes mellitus has a prominent role in causing Myocardial Infarction to human in most of the cases. Besides that the relevant parameters such as body mass index, fast blood sugar, varying hemoglobin, high density lipids, low density lipids, triglycerides, systolic blood pressure, diastolic blood pressure, and nicotine usage, etc seems to influence the occurrence of Myocardial Infarction were significantly showed by CHAID classifiers.

**Keywords**: Myocardial Infarction, Diabetics mellitus, CHAID, Data Mining.

## 1. INTRODUCTION

Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data [1]. Traditionally, the mined information is represented as a model of the semantic structure of the dataset. It might be possible to employ the model in the prediction and classification of new data [2].

Classification is one of the important decision making tasks for many real world problems. Classification will be used when an object needs to be classified into a predefined class or group based on attributes of that object. There are many real world applications that can be categorized as classification problems such as weather forecast, credit risk evaluation, medical diagnosis, bankruptcy prediction, speech recognition, handwritten character recognition, and quality control [3]. Generally, there are two types of classification problems: binary problem and multiclass problem. While a binary problem is a situation in which an outcome of prediction has to be determined with a decision of Yes or No, a multiple classification problem is a condition in which a predicted result is determined as multiple outcomes [4].

Researchers have long known that people who have diabetes mellitus (MI) are at very high risk of developing heart diseases. Cardiovascular disease (CVD) is the major cause of the morbidity and mortality associated with diabetes. A two to three fold incidence of CVD occurs in both type 1 and type 2 diabetics over that in age- and gender-matched non-diabetic persons [5]. Though many physicians seem to know these factors but many patients being ignorant of it.

It was reported that models constructed through data mining of inductive exercise were better in terms of prediction accuracy than those constructed through statistical measures with hypothetico-dedeuctive approach [6]. As data mining models have relatively higher degree of accuracy, we use such tools to develop predictive data mining model for Diabetes mellitus person [7].

In this paper, we propose a CHAID based method to diagnose heart disease for diabetic patients. It should be noted that the attributes used in our proposed method are those used for diagnosis of diabetes and are not direct indicators of heart disease.

## 2. BACKGROUND

Up to now, several studies have been reported that have focused on cardiovascular disease diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies of 77% or higher. Here are some examples:

A. Robert Detrano's experimental results showed correct classification accuracy of approximately 77% with logisticregression derived discriminant function [8].

B. Zheng Yao applied a new model called R-C4.5 which is based on C4.5 and improved the efficiency of attribution selection and partitioning models. An experiment showed that the rules created by R-C4.5s can give health care experts clear and useful explanations [9].

C. Resul Das introduced a methodology that uses SAS base software 9.13 for diagnosing heart disease. A neural networks ensemble method is at the center of this system [10].

D. Colombet et al. evaluated implementation and performance of CART and artificial neural networks comparatively with a LR model, in order to predict the risk of cardiovascular disease in a real database [11].

E. Engin Avci and Ibrahim Turkoglu study an intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases [12].

F. Imran Kurt, Mevlut Ture, A. Turhan Kurum compare performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease [13].

G. The John Gennari's CLASSIT conceptual clustering system achieved a 78.9% accuracy on the Cleveland database [14].

## 3. EXPERIMENTAL METHODOLOGY

### 3.1 Data collection and used variables

The data set used in this work are clinical data set collected from one of the leading diabetic research institute in Chennai and contain records of about 500 patients. The clinical data set specification provides concise, unambiguous definition for items related to diabetes. Data on the following attributes (Risk factors) were collected from 6073 diabetic subjects of MV Diabetics lab., Chennai, laying emphasis on the 267 subjects of MI.

The risk factors were *viz.*,

1. Body mass index (BMI)
2. Fast blood sugar (FBS)
3. Haemoglobin (HBAIC)
8. Diastolic blood pressure (BP-DIA)
9. Nicotine usage
10. Hypertension check (HT-CHECK)

4. High density lipids (HDL)
5. Low density lipids (LDL)
6. Triglycerides (TGL)
7. Systolic blood pressure (BP-SYS)
11. Gender / Sex
12. Cholesterol (CHO)
13. Diabetic-duration (DIA-DURATION).

Out of 6073, 257 subjects were characterized by Myocardial Infarction with due diagnostic features. Hence these 257 samples were given adequate importance for analysis. Each of these Myocardial Infarction subjects was closely monitored in terms of the thirteen risk factors which are instrumental for the cause of Myocardial Infarction in the diabetic patients. Among the risk factors Gender/Sex, Nicotine Usage and Hypertension check are categorical factors and Body mass index is calculated using height and weight parameters as under.

$$BMI = \frac{Weight}{(Height)^2} \; where\; height\; is\; measured\; in\; meters\; and\; weight\; in\; kgs$$

### 3.2. Tools and Techniques

Classification trees are widely used in different fields such as medicine, computer science, botany and psychology [15]. These trees readily lend themselves to being displayed graphically, helping to make them easier to interpret than they would be if only a strict numerical interpretation were possible.

CHAID (Chi Square Automatic Interaction Detection) [16], which is one of the classification tree algorithms, is the name given to one version of the Automatic Interaction Detector that has been developed for categorical variables. In fact, CHAID is a technique that recursively partitions (or splits) a population into separate and distinct segments. These segments, called nodes, are split in such a way that the variation of the response variable (categorical) is minimized within the segments and maximized among the segments. After the initial splitting of the population into two or more nodes (defined by values of an independent or predictor variable), the splitting process is repeated on each of the nodes. Each node is treated like a new sub-population. It is then split into two or more nodes (defined by the values of another predictor variable) such that the variation of the response variable is minimized within the nodes and maximized among the nodes. The splitting process is repeated until stopping rules are met i.e. when the class value in the partition is same or there is only one object in the partition. The output of CHAID prediction model is displayed in hierarchical tree-structured form, in which the root is the population, and the branches are the connecting segments such that the variation of the response variable is minimized within all the segments, and maximized among all the segments.

An essential step in CHAID prediction model construction is selecting the relevant features for classification [17]. The purpose of feature selection techniques helps in reduction of computation time and enhances the predictive accuracy of the model. Chi-square [14] is the common statistical test that measures divergence from the distribution expected if one assumes the feature occurrence is actually independent of the class value. Feature Selection via Pearson chi-square ($\chi^2$) test is a very commonly used method [18] and it evaluates features individually by measuring their chi-squared statistic with respect to the classes.

The present investigation used data mining as a tool with CHAID classification tree as a technique to design the Myocardial Infarction prediction model. Filtered feature selection technique [18] was used to select the best subset of variables on the basis of the values of chi-square measures.

## 4. RESULTS

In the present study, those features whose chi-square values were greater than 5 were given due considerations and the highly influencing variables with high chi-square values have been shown in Table 1. These features were used for the CHAID prediction model construction. For both variable selection and CHAID prediction model construction, we have used SPSS (version 16.0).

**Table 1: High Potential Variables**

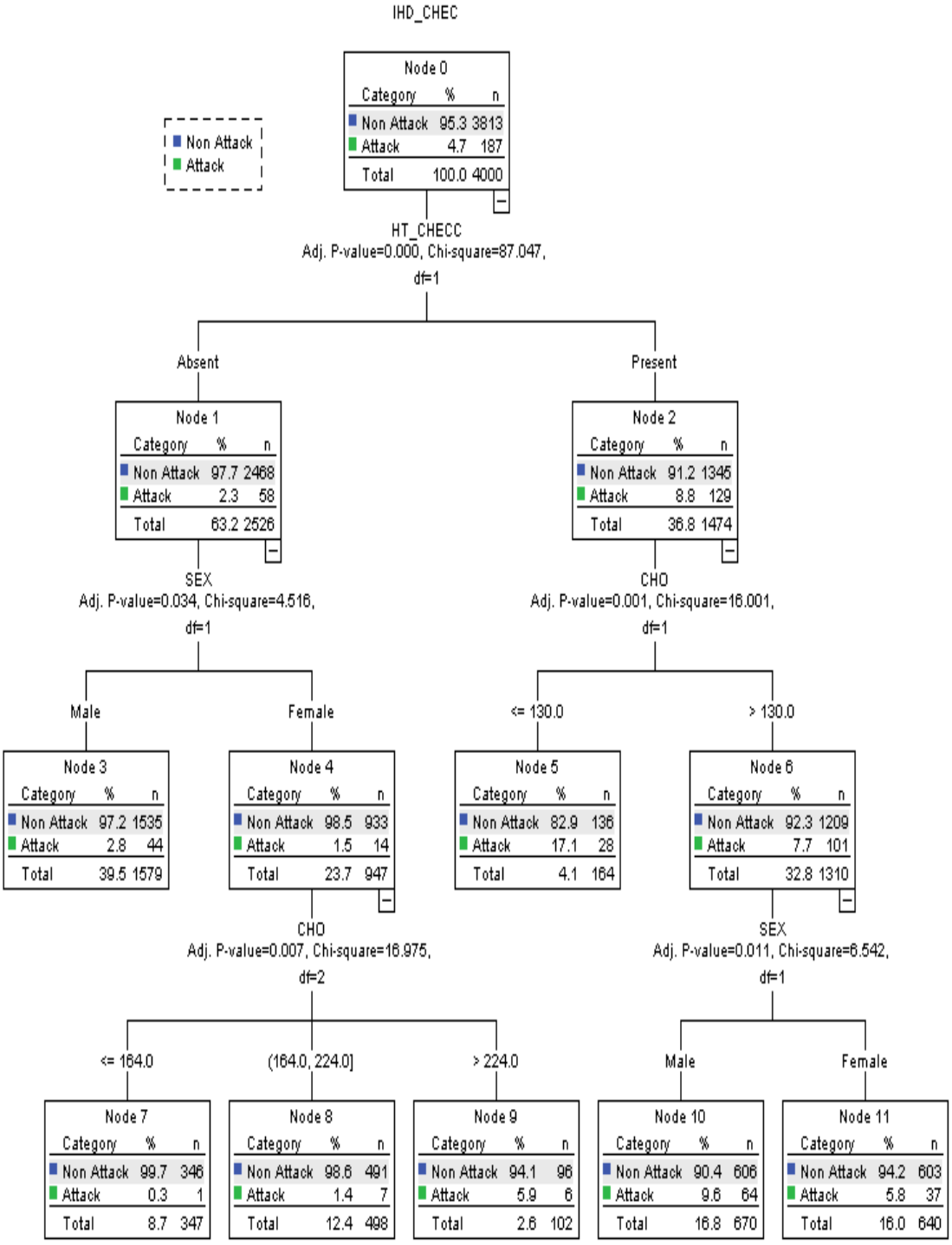
| Name of the Variable | Chi-Square Values |
|---|---|
| HT-CHECK | 87.047 |
| CHO | 16.975 |
| SEX | 6.542 |

A tree-based CHAID prediction model for the Myocardial Infarction was constructed using CHIAD algorithm that is shown in Fig. 1. Each node in Fig. 1 contains the details of *node-id* (ID), *number of data objects* (N), and *the possible outcome of the class variables (Attack, NON-Attack)* with high probability. There are 7 terminal nodes, resulting from 7 *if-then* conditions to predict the Attack and the Non-Attack of the Myocardial Infarction.

The tree starts with the top decision node (ID=0) with 4000 instances of the data set consisting of 3813 Non Attacks and187 Attacks. The whole data set is divided into two partitions based on the values of splitting predictive variable HT-CHECK ("*Absent*", "*Present*"). The left most node (ID=1) with 2526 *Absent of Hyper Tension* (HT-CHECK) out of which 2468 are Non Attack, 58 are Attack shows that the very less percentage (2.3%) of possibilities for Myocardial Infarction, whereas the right most node (ID=2) contains 1474 *present of Hyper Tension* (HT-CHECK) out of which 1345 are Non Attack, 120 are Attack shows that a relatively high percentage (8.8%) of possibilities for Myocardial Infarction.

The leftmost node (ID=1) containing 2526 instances is further split on the basis of the value of predictor variable-SEX (*Male, Female*), resulting in two more nodes (node ID=3 and node ID=4) and so on. Similarly the rightmost node (ID=2) containing 1474 instances is further split based on the predictor variable- CHO (*<=130,*>130) that results in two more nodes (node ID=5 and node ID=6) each containing 164(136-Non Attack, 28 Attack), 1310(1209-Non Attack, 101 Attack) instances respectively. The splitting process is explored on both sides further, until either split does not help to improve the predictive accuracy or a node contains instances which are less than the pre-defined size.

**Fig. 1 CHAID Tree Classification Model**



**Model Summary**

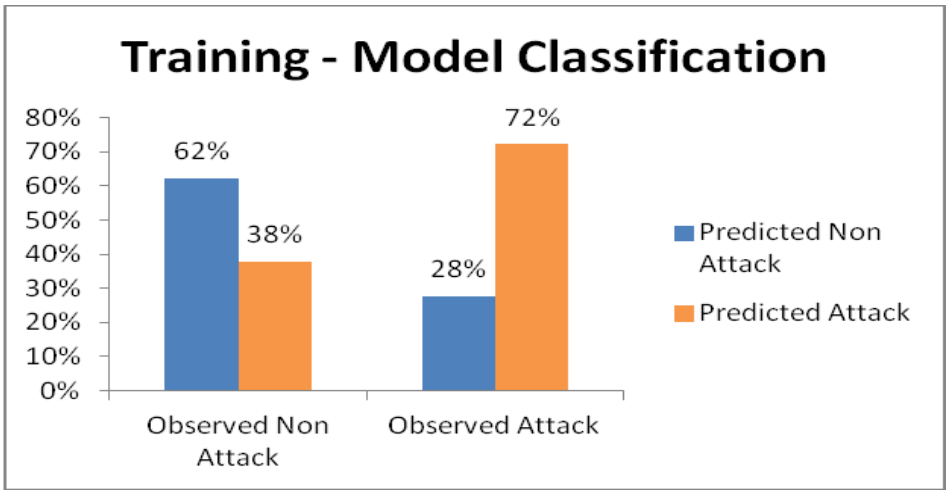
| | | |
|---|---|---|
| Specifications | Growing Method | CHAID |
| | Dependent Variable | IHD_CHEC |
| | Independent Variables | BMI, CHO, FBS, HBA1C, HDL, LDL, TGL, DIA_YEAR, HT_CHECC, SEX, NICOT, BP_SYS, BP_DIA |
| | Validation | Split Sample |
| | Maximum Tree Depth | 3 |
| | Minimum Cases in Parent Node | 200 |
| | Minimum Cases in Child Node | 100 |
| Results | Independent Variables Included | HT_CHECC, SEX, CHO |
| | Number of Nodes | 12 |
| | Number of Terminal Nodes | 7 |
| | Depth | 3 |

**Model Performance**

**Table 2: Performance of Training Data Set**

| | **OBSERVED** | | |
|---|---|---|---|
| **PREDICTED** | Non Attack | Attack | Grand Total |
| Non Attack | 2372 | 52 | 2424 |
| Attack | 1441 | 135 | 1576 |
| Grand Total | 3813 | 187 | 4000 |

**Chart 1: Training –Model Classification**



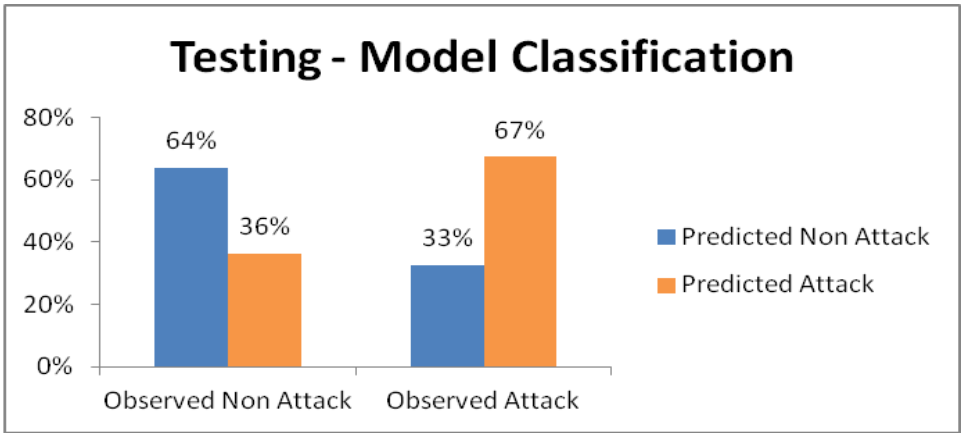

The Constructed CHAID Decision Tree is able to predict 72% of the Attack cases.

It indicates that the CHAID model could correctly classify 135 among 187 attack cases. And correctly classify 2372 among 3813 Non Attack cases.

**Table 3: Performance of Testing Data Set**

| | **OBSERVED** | | |
|---|---|---|---|
| **PREDICTED** | **Non Attack** | **Attack** | Grand Total |
| Non Attack | 606 | 16 | 622 |
| Attack | 345 | 33 | 378 |
| Grand Total | 951 | 49 | 1000 |

**Chart 2: Testing –Model Classification**




The Constructed CHAID Decision Tree is able to predict 67% of the Attack cases and hence a better medical treatment can be provided for these patients.

It indicates that the CHAID model could correctly classify 33 records among 49 records for attack cases. And correctly classify 606 records among 951 Non-Attack cases.

### 4.1. Rules Extracted from CHAID Tree

The classification rules can be generated by following the path from each terminal node to root node. Pruning technique was executed by removing nodes with less than desired number of objects.

**Table 4: Rule Set generated by CHAID Tree**

| **Rules for Predict** the Myocardial Infarction |
|---|
| IF HT_CHECC = '*Absent*' and SEX = '*Male*' THEN Attack =’2.8%’, Non-Attack=’97.2%’. |
| IF HT_CHECC = '*Absent*' and SEX = '*Female*' and CHO<=164 THEN Attack =’0.3%’, Non-Attack=’99.7%’. |
| IF HT_CHECC = '*Absent*' and SEX = '*Female*' and CHO BETWEEN 164-224 THEN Attack =’1.4%’, Non-Attack=’98.6%’. |
| IF HT_CHECC = '*Absent*' and SEX = '*Female*' and CHO>224 THEN Attack =’5.9%’,Non-Attack=’94.1%’. |
| IF HT_CHECC = '*Present*' and CHO<=130 THEN Attack =’17.1%’, Non-Attack=’82.9%’. |

| |
|---|
| IF HT_CHECC = '*Present*' and CHO>130 and SEX = '*Male*' THEN Attack =’9.6%’,Non-Attack=’90.4%’. |
| IF HT_CHECC = '*Present*' and CHO>130 and SEX = '*Female*' THEN Attack =’5.8%’,Non-Attack=’94.2%’. |

From the rule set generated by CHAID Tree, it’s clearly evident that Hyper Tension when present and cholesterol level less than or equal to 130 then the probability of attack becomes 17.1% and when the Hyper Tension is Absent in a Female and cholesterol level higher than 224 then the chances of rises by 5.9%. In an another case where Hyper Tension is present with Cholesterol level less than 130 increases the risk of attack by 9.6% for Male and 5.8% for Female.

## 5. CONCLUSIONS

The CHAID prediction model constructed from the study is useful to analyze the interrelation between variables that are used to predict the Myocardial Infarction. The risk factors like Hypertension check (HT-CHECK), Cholesterol (CHO), Gender / Sex are strongly influence the Myocardial Infarction.
The CHAID Prediction Model can be deployed by the doctors during their practice for better diagnosis.

## 6. ACKNOWLEDGEMENTS

We are grateful to **Dr. D. Sudarsanam,** Associate Professor, Department of Advanced Zoology, Loyola College, Chennai and **Prof. S. Chandrasekaran**, for providing an access to medical diabetic data and for his involvement in this domain. And also I would like to thank **Prof. M. Nester Jeyakumar, Prof. Sluvai Antony** and **Dr. M. Raja,** Loyola College, Chennai for their valuable suggestions.

## REFERENCES

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

[2] Sally Jo Cunningham and Geoffrey Holmes, "Developing innovative applications in agriculture using data mining", In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, 1999.

[3] G. P. Zhang, "Neural networks for classification: a survey," *Systems,Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 30, pp. 451-462, 2000.

[4] P. Kraipeerapun, "Neural network classification based on quantification of uncertainty," Murdoch University, 2008.

[5] Jennifer B. Marks, Philip Raskin *March-April 2000, Pages 108-115* Cardiovascular risk in diabetes: A brief review, *Journal of Diabetes and its Complications*, *Volume 14, Issue 2*.

[6] S. Ganesh, “Data Mining: Should it be Included in the Statistics Curriculum?” The 6th international conference on teaching statistics (ICOTS-6), Cape Town, South Africa, 2002.

[7] Kelling, D.G. and J.A. Wentworth et al., 1997, Diabetes mellitus. Using a database to implement a systematic management program. NC.Med.J.,58:368-371.

[8] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[9] Yao, Z.; Lei, L.; Yin, J., "R-C4.5 Decision tree model and its applications to health care dataset". Proceedings of International Conference on Services Systems and Services Management 2005, pp. 1099-1103.

[10] Das, R.; Abdulkadir, S. (2008). "Effective diagnosis of heart disease through neural networks ensembles". Elsevier, 2008.

[11] Colombet, I.; Ruelland, A.; Chatellier, G.; Gueyffier, F. (2000). "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression". Proceedings of AMIA Symp 2000, p 156-160.

[12] Avci, E.; Turkoglu, I., "An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases". Journal of Expert Systems with Application, Vol. 2, No. 1, 2009, pp. 2873-2878.

[13] Kurt, I.; Ture, M.; Turhan, A., "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". Journal of Expert Systems with Application, Vol. 3, 2008, pp. 366-374.

[14] Gennari, J., "Models of incremental concept formation" Journal of Artificial Intelligence, Vol. 1, 1989, pp. 11-61.

[15] H. A. Camdeviren, A. C. Yazici, Z. Akkus, R. Bugdayci, and M. A. Sungur, "Comparison of Logistic Regression Model and Classification Tree: An Application to Postpartum Depression Data", Expert Systems with Applications, Vol. 32, No. 4, 2007, pp. 987-994.

[16] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data", Applied Statistic, Vol. 29, 1980, pp. 119-127.

[17] I. H. Witten, and E. Frank, Data Mining – Practical Machine Learning Tools and Techniques (2nd ed.), San Francisco, CA: Morgan Kaufmann Publisher, 2005.

[18] H. Liu, and R. Setiono, "Chi-square: Feature Selection and Discretization of Numeric Attributes", Proceedings of IEEE 7th International Conference on Tools with Artificial Intelligence, Vol. 338, No. 391, 1995.