Engineering Sciences International Research Journal
ISSN (Print) : 2330 – 4338 ; Volume 1 Issue 1 ǁ February. 2013 ǁ

# An Integrated Approach for Predicting the Contribution of Sovereign Dynamics

**A.S. Aneeshkumar[1], C. Jothi Venkateswaran[2]**

***Abstract:*** The largest gland of the body, liver is seen in the upper right hand side of the abdomen, mostly behind the rib cage and its weight in an adult is around to three pounds. The vital functions performed by the liver are purifying the blood by changing harmful chemicals into harmless ones, production of many important substances like proteins and stores many sugars, fats and vitamins until they are needed elsewhere in the body for the physical activities.

Liver disorder is one of the major reliant disease and in recent years which rapidly increasing in developing countries as well as developed countries. Any other illnesses and physical conditions can also affect the functions of liver in addition to its own predicament. But yet the prevalence of risk factor for liver disease is not well characterized in most of the countries. So the potential study, which we carried out in this paper, is the analysis of the association between the psychodynamic- behavioural factors and liver syndrome.

***Keywords:*** Data mining, regression based classification.

## 1. INTRODUCTION

Data mining techniques are applied in huge database to focus on automatic exploration and analysis of large quantities of multidimensional data to extract meaningful patterns and rules [1]. For its approach formation data mining uses the availability of separate and integrated supporting fields like machine learning, artificial Intelligence, decision analysis and statistical models. The vast implementation of combined statistical and data mining task is applied in public and private sectors for the analysis and prediction of data. Data mining scopes are usually partitioned into predictive and descriptive areas with application specific changes pertaining to the requirements of the problem [2]. Predicting the data with the help of previously known values from some other data is done by predictive model where identification of models is made by descriptive model [3, 4].

In other life slaughter diseases like cardiovascular or kidney related nuisance, cancer and stroke the exact causing factors and variations of attack may easily calculated. Unlike this, liver allied muddle will not identify in the early stages without the combination of other defecting factors. Most patients are realizing their liver disorder, after reaching to the hospital for some other diagnosis. There are no preliminary symptoms of serious physical changes in the beginning stage. Liver is considered to be two to three percentage of body's mass. Dissimilar other organs like kidney, heart, stomach, liver does not carry with a single function. This largest gland has more than 140 functions [5]. Some of them are deactivation of toxic drugs, production of amino acids to build strong muscles, regulation of energy throughout the body, generation of bile required for digestion, piling up of minerals and vitamins, supporting for blood clotting, sustaining of hormonal balance and processing of sedatives [6]. Liver tissue is composed with thousands of lobules and each of them is made up of hepatic cells, which are the basic metabolic cells of the liver [7].

Various disorders can be affected liver with various depths; a few of them are cryptogenic hepatitis, drug toxity, primary and secondary biliary cirrhosis, autoimmune hepatitis, alcohol-related liver disease, primary sclerosing cholangitis, haemochromatosis, amyloid and glycogen storage diseases [8]. In accurate disorder, it will affect physical system and nervous system. Fatty liver or steatorrhoeic hepatitis is a reversible condition where large vacuoles of triglyceride fat accumulate in liver cells via the process of steatosis, which can see in people with high alcoholic consumption and sometimes in people them never or occasionally, had alcohol. Liver cancer may see in cirrhosis or certain type of hepatitis infected cases. Otherwise it may spread from other cancer affected organs. Cirrhosis refers to significant loss of cells [9]. The causing factor of various hepatitis are infection through food and water, contact with infected persons, sexual relations with infected persons, transfusion of blood, injected drug usage and diffusion from mother to the infant child [10].

## 2. TERM DEPICTION

Now a day, most health organizations are collecting the details of patients with the record of doctor's observation, symptoms, physical examination details, past history of such diseases or other important treatments and operations, blood reports, other image based reports, prescribed treatments and

improvements of that in a computer based storage device. So it helps to promote data analysis and modeling, and which lead to knowledge generation. In this study we collected 170 data from July 2011 through July 2012, with seven dependent factors of human life, where each field having more than one value and a class of diagnosis result.

**Table 1. List of independent factors**

| | |
|---|---|
| Job | 1,2,3,4 |
| Food | 1,2,3 |
| Drug usage | 1, 2,3,4,5 |
| Education | 1,2,3 |
| Exercise | 1,2,3 |
| Marital Status | 1,2 |
| Age | 1, 2,3,4,5 |
| Diagnosis | YES, NO |

According to Table 1, we are giving various numeric values for each attributes. For job we given four values, which are 1 for grade1, that means their monthly income is above fifty thousand. 2 represent the middle grade people, those earnings between ten thousand and fifty thousand, then grade 3 is for below ten thousand. In food 1: highly nutritious, 2: normal nutritious, 3: less nutritious and for drug usage 1: used earlier, 2 shows regular consumption now, 3 means weekly once now, 4: monthly once and 5 is occasional usage. Education- 1: highly educated, 2: normal education, 3: low education. Then in case of exercise 1 for regular, 2 for irregular, 3 for no and in case of marital status, 1 is for married and 0 is for unmarried. Finally for age: 1denotes age<15, 2=15< age<30, 3= 30< age<45, 4=45< age<60, 5= age>60.

## 3. STATISTICAL DATA MINING (SDM) TASKS

SDM extensively applied in economics, social science and as well as in scientific fields like experiments of physics, engineering, manufacturing, psychology and medicine [10]. Data Mining is the process of extracting hidden knowledge from the large data repositories [11]. For the extraction of hidden information and finding of relationship, we sometimes use statistical methods, which help to predict group membership for the instances of data. However, there are many well-established statistical techniques for data analysis, particularly for numeric data. Most of them are regression, generalized linear models, mixed effect models, factor analysis, discriminates analysis, time-series analysis, survival analysis and quality control.

## 4. ADOPTED METHODOLOGY:

***4.1 Regression:*** The regression methods are used for the prediction of response variable from one or more independent variables [12], where the variables are numeric. In other words we can say it as regression allows forecasting future values on the basis of past values. The strength of these variables can be evaluated by multivariate regression [13]. There are various forms of regression, such as linear, multiple, weighed, polynomial, nonparametric and robust [14].

***4.2 Logistic Regression:*** Logistic regression is used to model the relationship between a binary response variable and one or more predictor variables, which may be either discrete or continuous. Various outcom for a single data variable is common in medical applications. The probability modelling of the event occurrences as a function of linear set of predictor's variable is referred as logistic regression model [15, 16]. The distinguishing feature of logistic regression model is the binary or dichotomous. Probability of the outcome variable $Y$ is represented in logistic regression as,

$$P(Y) = \frac{1}{X} = \pi(\mathrm{X})$$

Where, $X = (x_1, x_2, \ldots\ldots\ldots\ldots x_k)$ is the collection of predicted variables.

$$\pi(X) = \frac{\exp\,(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ldots\ldots + \beta_k x_k)}{1 + \exp\,(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ldots\ldots + \beta_k x_k)}$$

$\beta_0$ is a common constant and where $\beta_1\ to\ \beta_k$ is the symbolic representation of model parameters or constants for each attributes.

## 5. SIGNIFICANCE OF THE MODEL

Determining the significance or goodness of fit is a statistical measure for explain how well a model fits an observation set. The model is more significant, if the difference is larger and the evidence is greater.

***5.1. Hosmer-Lemeshow Statistic:*** One of the most straightforward measures of a goodness of fit is the Hosmer-Lemeshow (HL) statistic. It groups the predictions of a logistic regression model rather than the model's predictor variable data, which is the Pearson statistic's approach.

$$HL = \sum_{j=1}^{k} \frac{(O_j - n_j\,\bar{\pi}_j\,)^2}{n_j\,\bar{\pi}_j(1 - \bar{\pi}_j)}$$

Where, $O_j$ is the number of positive observations in group $j, \bar{\pi}_j$ is the model's average predicted value in group $j$, and $n_j$ is the size of the group. The HL statistic follows a chi-squared distribution with $k$ - 2 degrees of freedom.

***5.2. Parameter Significance (Wald Test):***

Wald statistic test is basically identical to the t-test in linear regression and the value of the regression coefficient divided by its associated standard error. According to Wald test value

all the independent variables are highly significant predictors for outcome and which is defined as,

$$W_j = \left(\frac{\mu_j}{SE_{\beta_j}}\right)^2$$

Total Number of Instances: 170

$$Where, j = 1, 2, \ldots\ldots\ldots\ldots, k$$

Correctly Classified Instances: 123 (72%)
Incorrectly Classified Instances: 47(28%)
Kappa statistic: 0.2973
Mean absolute error: 0.3891
Root mean squared error: 0.4411
Relative absolute error: 87.172 %
Root relative squared error: 93.432%
Coverage of cases (0.95 levels): 100 %
Mean rel. region size (0.95): 99.76 %

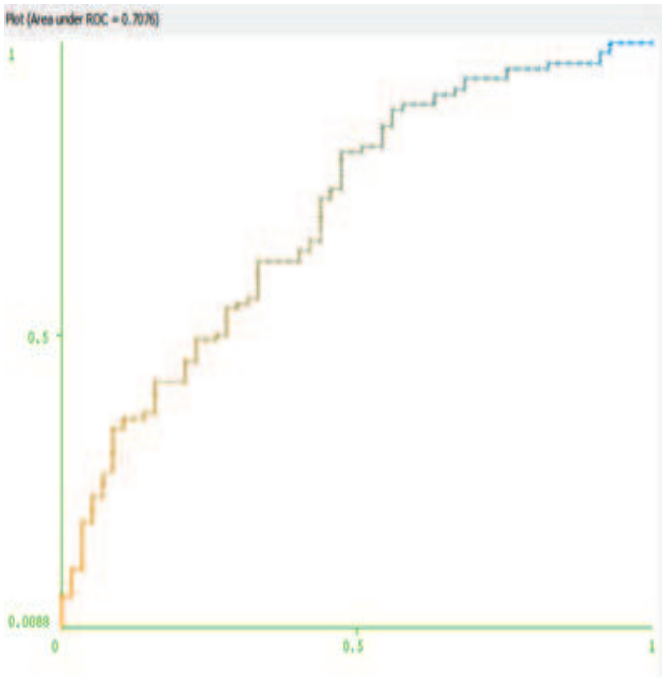


**Fig. 1. ROC curve for Liver diagnosis**

**Table2. Logistic Regression with ridge Parameter of 1.0E-8 coefficients and ratio**

| Variable | Coefficients | Ratio | Significance |
|---|---|---|---|
| Job | 0.0226 | 1.0229 | 0.782 |
| Food | 0.8175 | 2.2648 | 0.165 |
| Drug usage | 0.2232 | 1.25 | 0.5 |
| Education | 0.5917 | 1.8071 | 0.483 |
| Exercise | - 0.4717 | 0.6239 | 0.405 |
| Marital Status | 0.1331 | 1.1424 | 0.824 |
| Age | 0.0392 | 1.0399 | 0.976 |

**Table 3. Other related results**

| Item | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|---|---|---|---|---|---|---|
| Patients | 0.912 | 0.649 | 0.736 | 0.912 | 0.814 | 0.708 |
| Weighted Avg. of both classes | 0.724 | 0.461 | 0.713 | 0.724 | 0.695 | 0.708 |

## 6. DISCUSSION

A total of 170 data are used as training set for building this model, out of this 113 cases are suffering from liver diseases and 57 are not, where the correlation coefficient, ratio and significance of each attribute for diseased cases is shown in Table2. As per this result, all the variables having positive correlation while exercise having negative correlation coefficient value of (-0.4717). According to other results we can see the root mean squared error and absolute error rates. The accuracy of this method is 72.35 % and incorrect classification accuracy is 27.65 %.

The quality of the classification is always recognized with the help of confusion matrix, because other given measures indicate only the total number of correct classifications. In medical diagnosis, it is necessary for a more detailed accuracy analysis including the number of FP, FN, TP and TN [17]. In Table 3, we are having TP rate and FP rate, which defined as,

$$TP\ Rate = \frac{TP}{TP + FP}$$

$$and$$

$$FP\ Rate = \frac{FP}{FP + TN}$$

Where TN (True Negative) predicts negative result as negative, FP (False Positive) predicts a negative person as positive, FN (False Negative) predicts positive as negative and TP (True Positive) predicts positive result as positive [18, 19]. However we need the accurate value of this. So the confusion matrix (CM) in needed to see the cross-classification of the predicted class against the true class as,

$$CM = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} = \begin{bmatrix} 103 & 10 \\ 37 & 20 \end{bmatrix}$$

The rows represent the true classes and columns represent the predicted classes. The ROC curve for the classifier is shown in Figure 1**.** ROC curves can be used for estimation of parametric or non-parametric assumptions [20]. Plotting the ROC curve is a popular way of displaying the discriminatory accuracy of a diagnostic test for detecting whether or not a patient has a disease or condition [21]. That is a person is

assessed as diseased (positive) or healthy (negative) depending on whether the corresponding marker value is greater than or less than or equal to a given threshold value [22].

## 7. CONCLUSION

In this paper we developed a model to identify liver infected cases from various independent factors. So this methodology can be applicable in other sectors also for correlation analysis and prediction. In further we will extend this work for obtain improved and accurate results using other data mining techniques. According to this analysis, the independent fields like profession, diet plans, drug usage, educational differentiation, age and marital status also having positive correlation with liver disorder.

## 8. REFERENCES


[1] Duke Hyun Choi, Byeong Seok Ahn and Soung Hioe kim, " Prioritization of association rule in data mining: Multiple criteria decision approach", Expert System with Applications, ELSEVIER(0957-4174), 2005.

[2] Huda Yasin, Tahseen A. Jilani and Madiha Danish, "Hepatitis-C Classification using Data Mining Techniques", International Journal of Computer Applications(0975-8887), Volume 24-No.3, June 2011.

[3] Dunham M.H. and Sridhar S., "data mining: Introductory and Advanced topics, Pearson Education, 2006.

[4] Larose D.T., "Data Mining Methods and Models", Jphn Wiley and sons, 2006.

[5] World Health Organization, Hepatitis C (Fact Sheet Page No.164), Geneva- 2000.

[6] World Health Organization, Hepatitis C Global Prevalence, Weekly Epidemiological Record, 74, 1999, Page 421-427.

[7] P.rajeswari and G. Sophia Reena, "Analysis of Liver Disorder Using Data Mining Algorith", Global Journal of Computer Science and Technology, Vol.10 Issue 14(Ver.1.0) November 2010.

[8] Maeve M. Skelly, Peter D. James and Stephen D. Ryder, "Findings on liver biopsy to investigate abnormal liver function tests in the absence of diagnostic serology", Journal of Hepatology, ELSEVIER-2001.

[9] P.Rajeswari and G.Sophia reena, "Analysis of Liver Disorder using Data Miniing Algorithm", Global Journal of Computer Science and Technology, Vol.10 issue.14(Ver. 10) November 2010 Page. 48.

[10] http://www.puristat.com/liver-cleansing/liver-disease-symptoms.aspx

[11] S.Vijayarani and M.Divya, "An Efficient Algorithm for Generating Classification Rules", IJCST Vol.2, Issue 4, Oct-Dec 2011.

[12] Rogue Wave Software online documentation, "Logistic regression", http//:www.3.3 Logistic Regression.html.

[13] Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III, pp. 548 – 555, Springer-Verlag.

[14] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Published by Elsevier, second edition – 2006.

[15] M. H. Dunham and S. Sridhar, Data Mining: Introductory and Advanced topics, Pearson Education 2006, chapter 1, chapter 3,chapter 4.

[16] Sharareh R. Niakan Kalhori, Mahshid Nasehi and Xiao-Jun Zeng, "A Logistic Regression Model to Predict High Risk Patients to Fail in Tuberculosis Treatment Course Completion", IAENG International Journal of Applied Mathematics, 40:2, IJAM_40_2_08.

[17] Brain Leke-Betechuoh, Tshhilidzi Marwala, Taryn Tim and Monica Lagazio, " Prediction of HIV Status from Demographic Data using Neural Networks", IEEE, December 16, 2005.

[18] Tahseen A.Jilani, Huda Yasin, Madiha Yasin and Cemal Ardil, "Acute Coronary Syndrome Prediction using Data Mining Techniques- An Application", World Academy of Science, Engineering and Technology 59 2009.

[19] Lee, W., S. J. Stolfo, and K. W. Mok, " Mining in a data-flow environment: Experience in network intrusion detection," In S. Chaudhuri and D. Madigan (Eds.), Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, CA, pp. 114124. ACM,12-15 August 1999.

[20] Shapiro DE. The interpretation of diagnostic tests. Statistical Methods in Medical Research 1999; 8:113–134.

[21] David Faraggi and Benjamin Reiser, "Estimation of the area under the ROC curve", STATISTICS IN MEDICINE, Statist. Med. 2002; 21:3093–3106 (DOI: 10.1002/ sim.1228).

[22] Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver-operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in Medicine 1997;16;2143–2156.


* * *


*[1]Research Scholar, PG and Research Department of Computer Science,*
*Presidency College (Autonomous), Chennai, India – 600 005.*
*aneesh_kumar777@yahoo.com*

*[2]Research supervisor & Dean, Department of Computer Science & Applications,*
*Presidency College (Autonomous), Chennai, India – 600 005.*
*jothivenkateswaran@yahoo.co.in*