



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Artificial Intelligence Approach for Disease Diagnosis and Treatment

S. Vaishnavi

PG Scholar, ME- Department of CSE, PSR Engineering College, Sivakasi, TamilNadu, India.

ABSTRACT: Generally, Data mining plays an important role in prediction of diseases in health care industry. The availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. Medical data are an ever-growing source of information generated from the hospitals in the form of patient records. When mined properly, the information hidden in these records is a huge resource bank for medical research. In this Project, the aim is Medical decision is a highly specialized and challenging job due to various factors, especially in case of diseases that show similar symptoms, or in case of rare diseases. It is a major topic of artificial intelligence in medicine. A Diagnosis Decision Support Systems(DDSS) would take the patients data and propose a set of appropriate Prediction. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease and use the medical profiles such as age, Blood Pressure and Blood Sugar it can predict the likelihood of patients getting a heart disease. Classification algorithm that has been used with the number of attributes for prediction. Web based questionnaire application can serve a training tool to diagnose the patients with disease. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

KEYWORDS: Data mining, knowledge discovery, medical decision support system, medical profiles, Naive bayes, Heart Disease Prediction System(HDPS).

I. INTRODUCTION

In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease. As a result of this people go to healthcare practitioners but the prediction made by them is not 100% accurate [1].

Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems [2].

II. RELATED WORK

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [3]. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions .The main objective of this research is to develop a prototype Heart Disease Prediction System (HDPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network [4]. So it provides effective treatments, it also helps to reduce treatment costs and also enhances visualization and ease of interpretation.

The main objective of this research is to develop a prototype Health Care Prediction System using, Naive Bayes .The System can discover and extract hidden knowledge associated with diseases (heart attack, cancer and diabetes) from a historical heart disease database. It can answer complex queries for diagnosing disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results in tabular and PDF forms.

III. PROPOSED ALGORITHM

A. Data Sources:

Clinical databases have accumulated large quantities of information about patients and their medical conditions. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease is the major cause of casualties in the world. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease kills one person every 34 seconds in the United States.

Record set with medical attributes was obtained from the UCI Repository With the help of the dataset, the patterns significant to the heart attack prediction are extracted.

The attribute "Diagnosis" is identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. "Patient Id" is used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

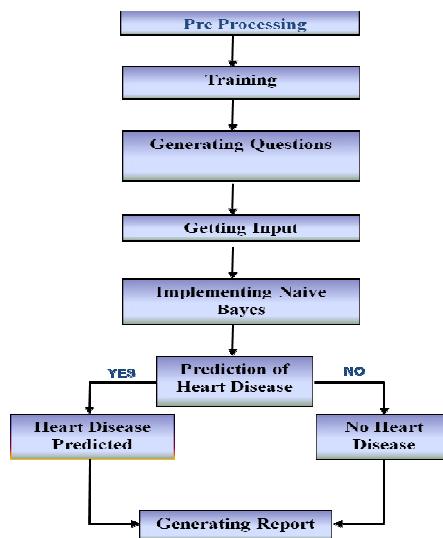


Figure:1 Block diagram for Heart disease diagnosis

B. Data set Generation:

Questionnaires have advantages over some other types of medical symptoms that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data [8]. However, such standardized answers may frustrate users. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them. Here our questionnaire is based on the attribute given in the data set, so the questionnaire contains:

Predictable attribute:

1. Diagnosis (value 0: <50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (has heart disease))

Key attribute:

1. Patient Id – Patient's identification number.

Input attributes:

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl).
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy).
5. Exang – exercise induced angina (value 1: yes; value 0: no).
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping).
7. CA – number of major vessels colored by floursopy (value 0 – 3).
8. Thal (value 3: normal; value 6: fixed defect; value 7:reversible defect)
9. Rest Blood Pressure (mm Hg on admission to the hospital).
10. Serum Cholesterol (mg/dl).
11. Thalach – maximum heart rate achieved.
12. Oldpeak – ST depression induced by exercise relative to rest.
13. Age in Year.
14. Height in cms.
15. Weight in Kgs.

C. Data set Analysis:

A total of 500 records with 15 medical attributes (factors) were obtained from the Heart Disease database lists the attributes. The records were split into two datasets such as training dataset (455 records) and testing dataset (454 records). To avoid bias, the records for each set were selected randomly.

In artificial intelligence or machine learning, a training set consists of an input vector and an answer vector, and is used together with a supervised learning method to *train* a knowledge database (e.g. a neural net or a naive bayes classifier) used by an AI machine.

In a dataset a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. After a model has been processed by using the training set, test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute to predict.

IV. SIMULATION AND RESULTS

A. Classifier

A classifier is a process of mapping from a (discrete or continuous) feature space X to a discrete set of labels Y . Here we are dealing about learning classifiers, and learning classifiers are divided into supervised and unsupervised learning classifiers [2]. The applications of classifiers are wide- ranging. They find use in medicine, finance, mobile phones, computer vision (face recognition, target tracking), voice recognition, data mining and uncountable other areas. An example is a classifier that accepts a person's details, such as age, marital status, home address and medical history and classifies the person with respect to the conditions of the project.

B. Naive Bayes

In probability theory, Bayes' theorem (often called Bayes' law after Thomas Bayes) relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations [2]. For example, a patient may be observed to have certain symptoms. Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation. A naive Bayes classifier is a term dealing with a simple probabilistic classification based on applying Bayes' theorem. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting [7]. Naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event.



ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers [4].

An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix [8].

C. Theorem

A conditional probability is the likelihood of some conclusion, C , given some evidence/observation, E , where a dependence relationship exists between C and E .

This probability is denoted as $P(C | E)$ where,

$$P(C | E) = \frac{P(E | C)P(C)}{P(E)}$$

The Bayesian Classifier is capable of calculating the most probable output depending on the input. It is possible to add new raw data at runtime and have a better probabilistic classifier. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

D. Bayesian interpretation

In the Bayesian (or epistemological) interpretation, probability measures a *degree of belief*. Bayes' theorem then links the degree of belief in a proposition before and after accounting for evidence [4]. For example, suppose somebody proposes that a biased coin is twice as likely to land heads as tails. Degree of belief in this might initially be 50%. The coin is then flipped a number of times to collect evidence. Belief may rise to 70% if the evidence supports the proposition [2].

For proposition A and evidence B ,

$P(A)$, the *prior*, is the initial degree of belief in A . $P(A | B)$, the *posterior*, is the degree of belief having accounted for B . $P(B | A) / P(B)$ represents the support B provides for A .

E. Example Classification

Sex classification - Classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size. A sample to be classified as a male or female. To determine which posterior is greater, male or female. For the classification as male the posterior is given by

| sex | height (feet) | weight (lbs) | foot size (inches) |
|--------|---------------|--------------|--------------------|
| male | 6 | 180 | 12 |
| male | 5.92 (5'11") | 190 | 11 |
| male | 5.58 (5'7") | 170 | 12 |
| male | 5.92 (5'11") | 165 | 10 |
| female | 5 | 100 | 6 |
| female | 5.5 (5'6") | 150 | 8 |
| female | 5.42 (5'5") | 130 | 7 |
| female | 5.75 (5'9") | 150 | 9 |

Table 1. Training Dataset

The classifier created from the training set using a Gaussian distribution assumption would be:



ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

| sex | mean (h) | variance (h) | mean (w) | variance (w) | mean (foot size) | variance (foot size) |
|--------|----------|--------------|----------|--------------|------------------|----------------------|
| male | 5.855 | 3.5033e-02 | 176.25 | 1.2292e+02 | 11.25 | 9.1667e-01 |
| female | 5.4175 | 9.7225e-02 | 132.5 | 5.5833e+02 | 7.5 | 1.6667e+00 |

Table 2. Values calculated by Gaussian distribution

Let's say we have equiprobable classes so $P(\text{male}) = P(\text{female}) = 0.5$. There was no identified reason for making this assumption so it may have been a bad idea. If we determine $P(C)$ based on frequency in the training set, we happen to get the same answer. Below is a sample to be classified as a male or female.

| sex | height (feet) | weight (lbs) | foot size (inches) |
|--------|---------------|--------------|--------------------|
| sample | 6 | 130 | 8 |

Table 3. Sample for classification

To determine which posterior is greater, male or female. For the classification as male, the posterior is given by:

$$\text{Posterior (male)} = \frac{p(\text{male})p\left(\frac{\text{height}}{\text{male}}\right)p\left(\frac{\text{weight}}{\text{male}}\right)p\left(\frac{\text{footsize}}{\text{male}}\right)}{\text{evidence}}$$

$$\text{Posterior (female)} = \frac{p(\text{female})p\left(\frac{\text{height}}{\text{female}}\right)p\left(\frac{\text{weight}}{\text{female}}\right)p\left(\frac{\text{footsize}}{\text{female}}\right)}{\text{evidence}}$$

The evidence (also termed normalizing constant) may be calculated since the sum of the posteriors equals one. It learns from the "evidence" by calculating the correlation between the target (dependent) and other (independent) variables. The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

The evidence may be ignored since it is a positive constant. (Normal distributions are always positive.) We now determine the sex of the sample. $P(\text{male}) = 0.5$

$$p(\text{height} | \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(6 - \mu)^2}{2\sigma^2}\right) \approx 1.5789$$

Where $\mu = 5.855$ and $\sigma^2 = 3.5033e - 02$ are the parameters of normal distribution which have been previously determined from the training set. Note that a value greater than 1 is OK here – it is a probability density rather the probability, because height is a continuous variable.

$$P(\text{weight} | \text{male}) = 5.9881e-06$$

$$P(\text{foot size} | \text{male}) = 1.3112e-3$$

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Posterior numerator (male) = their product = 6.1984e-09.

$$P(\text{female}) = 0.5$$

$$p(\text{height} | \text{female}) = 2.2346e-1$$

$$p(\text{weight} | \text{female}) = 1.6789e-2$$

$$p(\text{foot size} | \text{female}) = 2.8669e-1$$

$$\text{Posterior numerator (female)} = \text{their product} = 5.3778e-04$$

Since posterior numerator is greater in the female case, we predict the sample is female.

step:1

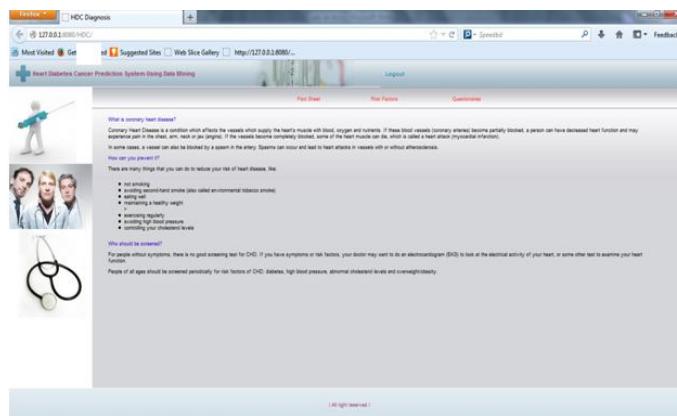


Fig. 1. CHD fact sheet and risk factors

step:2

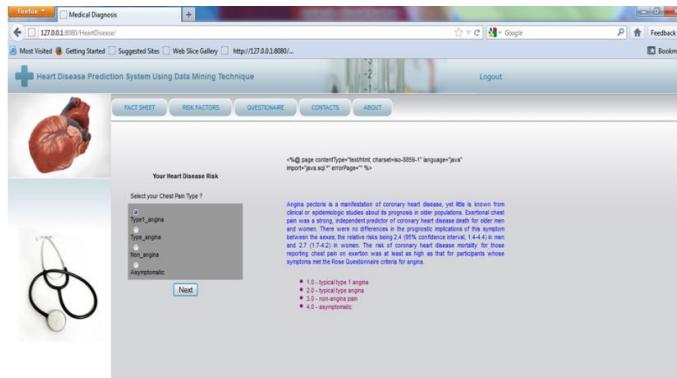


Fig. 2. Chest pain type in questionnaire

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

step:3

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|-------------------------|----|----------|---|----------|---------|-----|----|-----|------|------------|--------------|----------------|-----|-------------|-----|-------------|------|---------|-------------------------|-------|----|--------|-----|--------|----|--|--------|------------------|
| 1. User Information 1.1. Personal Detail <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 5px;"> <tr><td>Userid</td><td>24</td></tr> <tr><td>UserName</td><td>F</td></tr> <tr><td>phone No</td><td>4544444</td></tr> </table> 1.2. Report Time : Sat Oct 12 12:05:50 IST 2013 1.3. Entered Details: <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 5px;"> <tr><td>Age</td><td>60</td></tr> <tr><td>Sex</td><td>Male</td></tr> <tr><td>Chest Pain</td><td>TypeI_angina</td></tr> <tr><td>Blood Pressure</td><td>110</td></tr> <tr><td>Cholesterol</td><td>198</td></tr> <tr><td>Blood Sugar</td><td>>100</td></tr> <tr><td>Resteeg</td><td>Ventricular_hypertrophy</td></tr> <tr><td>Trich</td><td>89</td></tr> <tr><td>Eating</td><td>yes</td></tr> <tr><td>Obseak</td><td>70</td></tr> </table> | Userid | 24 | UserName | F | phone No | 4544444 | Age | 60 | Sex | Male | Chest Pain | TypeI_angina | Blood Pressure | 110 | Cholesterol | 198 | Blood Sugar | >100 | Resteeg | Ventricular_hypertrophy | Trich | 89 | Eating | yes | Obseak | 70 | 1.4. Prediction Result: <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 5px;"> <tr><td>Report</td><td>No heart disease</td></tr> </table> | Report | No heart disease |
| Userid | 24 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| UserName | F | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| phone No | 4544444 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Age | 60 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sex | Male | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Chest Pain | TypeI_angina | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Blood Pressure | 110 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Cholesterol | 198 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Blood Sugar | >100 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Resteeg | Ventricular_hypertrophy | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Trich | 89 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Eating | yes | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Obseak | 70 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Report | No heart disease | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Fig. 3 Report generation

step:4

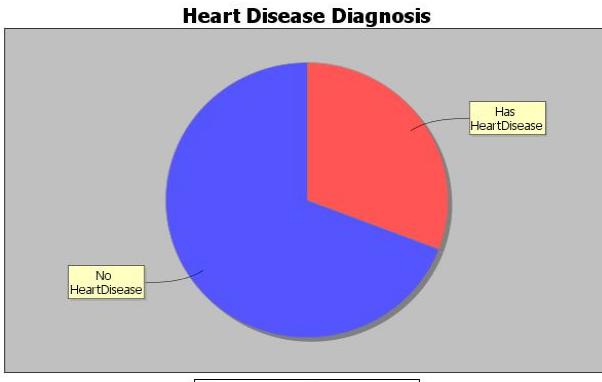


Fig. 4. Heart disease prediction chart

IV. CONCLUSION AND FUTURE WORK

Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. HDPS can be further enhanced and expanded. For, example it can incorporate other medical attributes besides the above list. It can also incorporate other data mining techniques. Continuous data can be used instead of just categorical data and the classifier separates lower risk patients from higher risk ones.

REFERENCES

- [1] M. Berlingero, F. B. F. Giannotti, and F. Turini, "Mining clinical data with a temporal dimension," in Proc. IEEE Int'Conf .Bioinf. Biomed, 2007.
- [2] Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

- [3] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success," Br.Med. 2005.
- [4] Shantakumar B. Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.3 1 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.
- [5] R. Carvalho, R. Isola, and A. Tripathy, "MediQuery—An automated decision support system," in Proc. 24th Int. Symp. Comput.-Based Med. Syst, 2011.
- [6] Nidhi Bhatla, Kiran Jyoti , "An Analysis of Heart Disease Prediction using Different Data Mining Techniques"; International Journal of Engineering Research & Technology (IERT) Vol. 1 Issue 8, ISSN: 2278-0181 , October 2012.
- [7] Walid Moudani , "Dynamic Features Selection for Heart Disease Classification"; World Academy of Science, Engineering and Technology 74 2013.
- [8] S. Vijayarani, S.Sudha , "Disease Prediction in Data Mining Technique"; International Journal of Computer Applications & Information Technology Vol. II, Issue I, (ISSN: 2278-7720), January 2013.
- [9] V. Sree Hari Rao and M. Naresh Kumar, "Novel Approaches for Predicting Risk Factors of Atherosclerosis," IEEE Trans. Inf.Technol. Biomed., vol. 17, no. 1, January 2013.
- [10] Leandro Pecchia, Paolo Melillo, and Marcello Bracale, "Remote Health Monitoring of Heart Failure With Data Mining via CART Method on HRV Features"; IEEE Trans., Biomedical Eng., VOL. 58, NO. 3, MARCH 2011.

BIOGRAPHY

S. Vaishnavi is a ME(CSE) Student in the Department of Computer Science and Engineering, PSR Engineering College, Sivakasi, TamilNadu, India. She received BE(CSE) degree in from Kalasalingam University, Krishnankoil, Tamil Nadu, India. Her research interests are Data Warehousing and Mining, Cloud Computing and Image Processing.