The Prediction of Myocardial Infarction for Heart Patients

[1]Swathi K, [2]Sai Shashikaa T R, [3]Swetha Adiga G , [4]Rabiya Farheen, [5]Vaishnavi K N

[1]Assistant Professor , [2,3,4,5]Student

[1,2,3,4,5]Department of Computer Science & Engineering

[1,2,3,4,5]K.S.Institute of Technology,Bengaluru – 560062 , Karnataka, India

Abstract : In this paper, by using data mining we can evaluate many patterns which will be use in future to make intelligent systems and decisions By data mining refers to various methods of identifying information or the adoption of solutions based on knowledge and data extraction of these data so that they can be used in various areas such as decision-making, the prediction value for the prediction and calculation. In our days the health industry has collected vast amounts of patient data, which, unfortunately, is not "produced" in order to give some hidden information, and thus to make effective decisions, which are connected with the base of the patient's data and are subject to data mining. This research work has developed a Decision Support in Heart Disease Prediction System (HDPS) using data mining modeling technique. Using of medical data, such as age, sex, blood pressure and blood sugar levels, chest pain, electrocardiogram, analyzes of different study patient, etc. graphics can predict the likelihood of the patient.

Keywords **- GUI PyQt5, MLP, Naive Bayes, SVM, Prediction, Accuracy.**

I. INTRODUCTION

Electronic health records (EHR) are rapidly ending up increasingly basic among human services offices. With expanded access to a lot of patient information, human services suppliers would now be able to improve the effectiveness and nature of their associations utilizing information mining. Since the 1990s, organizations have utilized information digging for things like credit scoring and misrepresentation discovery. Presently, various human services associations are additionally starting to see the potential advantages of information mining and prescient examination. In social insurance, information mining has demonstrated powerful in regions, for example, prescient prescription, client relationship the executives, location of misrepresentation and misuse, the executives of medicinal services and estimating the adequacy of specific medications. The motivation behind information mining, regardless of whether it's being utilized in human services or business, is to distinguish helpful and justifiable examples by dissecting substantial arrangements of information. These information designs help foresee industry or data patterns, and after that figure out what to do about them. Coronary illness is the greatest reason for death these days. Circulatory strain, cholesterol, beat rate are the significant purpose behind the coronary illness. Some non-modifiable variables are additionally there. For example, smoking, drinking likewise explanation behind coronary illness. The heart is a working arrangement of our human body. On the off chance that the capacity of heart isn't done legitimately implies, it will influence other human body part moreover. Some hazard factors of coronary illness are Family history, High circulatory strain, Cholesterol, Age, Poor eating regimen, Smoking. At the point when veins are overstretched, the hazard dimension of the veins are expanded. This prompts the circulatory strain. Pulse is ordinarily estimated as far as systolic and diastolic. Systolic demonstrates the weight in the supply routes when the heart muscle contracts and diastolic shows the weight in the veins when the heart muscle is in resting state. The dimension of lipids or fats expanded in the blood are causes the coronary illness. The lipids are in the supply routes consequently the veins become tight and blood stream is additionally turned out to be moderate. Age is the non-modifiable hazard factor which additionally an explanation behind coronary illness. Smoking is the explanation behind 40% of the passing of heart infections. Since it restrains the oxygen level in the blood then it harm and fix the veins.

II. RELATED WORK

Intelligent decision support systems are defined as interactive computer systems to help make decisions in the use of data sets and models to find problems, solve problems and make decisions [1]. Medical record attribute set has been obtained from the database Cleveland heart disease. With the help of a set of data patterns is vital for predicting heart attack removed. Records were divided equally into two data sets: training dataset and testing dataset [4].Naive Bayes algorithm is optimal, that is, has the minimum error probability. It calculates the precise probability of the hypothesis, and it is resistant to the noise at the input [5] data. If dependency is discovered between two values B_i and B_j of two different attributes, in this case the data are not considered as conditionally independent [5]. By using the Naïve Bayes method, possible attributes will be determined and probability of each attribute will be calculated. Then yes or no probability of each attribute will be computed, and depending on these results the information about risk will be returned [6].Artificial agents are part of the system that have own knowledge base and can communicate with main database and also agents will help to solve part of the main problem which showed up in the result of dividing several sub-problems of the diagnosis problem [7].

III. EXISTING SYSTEM

The expanding volume of social insurance information contained in Electronic Health Records (EHRs) has made many think about planning computerized clinical help and sickness recognition frameworks dependent on patient history and hazard factors. Various past investigations have endeavored to utilize tolerant research facility tests, determinations, and meds as methods for anticipating illness beginning. Such models have likewise been utilized to recognize possibly obscure hazard factors, frequently while at the same time improving affectability and explicitness of discovery. Various ongoing examinations have been effective in foreseeing ailment through different strategies, including bolster vector machines, calculated relapse, arbitrary timberlands, neural systems, and time arrangement demonstrating methods. Many have noticed that profound learning strategies have been especially effective for offering new understanding into the two information portrayal and finding in drug.

IV. PROPOSED SYSTEM

In this model, we are utilizing three unique information mining procedures in particular – Decision Tree calculation for grouping and furthermore Linear relapse. Alongside the characterizations the models are utilized to foresee traits, for example, age, sex, circulatory strain and glucose for odds of a patient getting coronary illness. We are likewise looking at both the strategies of characterization. By looking at it we can demonstrate which calculation is better over the other for characterization. Points of interest of Proposed System: 1. Better outcome precision is appeared. 2. Decreased time intricacy**.**

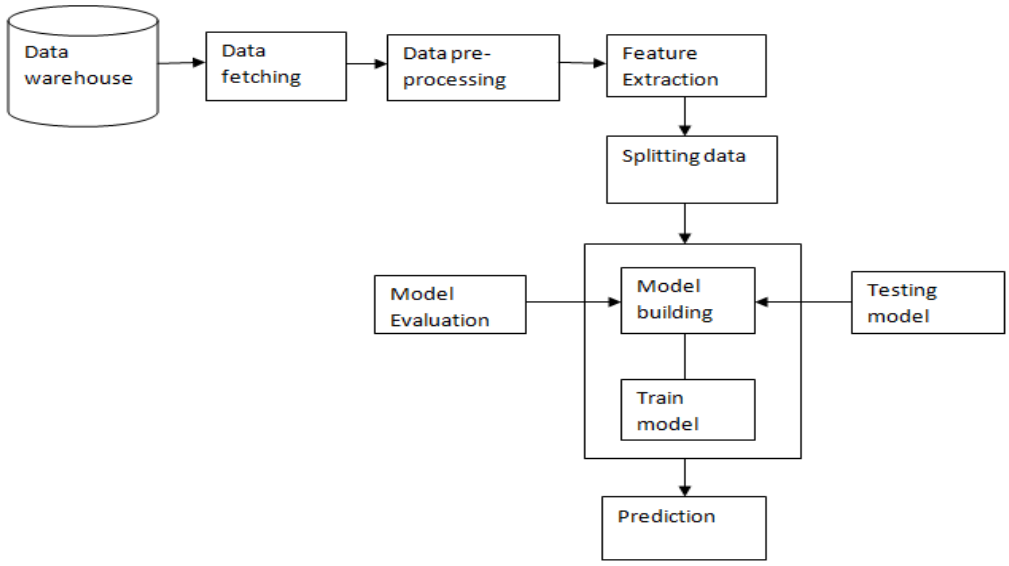

Fig 1: Proposed architecture

An information distribution center (DW) is a gathering of corporate data and information got from operational frameworks and outside information sources. An information stockroom is intended to help business choices by permitting information solidification, examination and detailing at various total dimensions. The information is gotten from the information stockroom. Information Preprocessing is a method that is utilized to change over the crude information into a perfect informational collection. At the end of the day, at whatever point the information is assembled from various sources it is gathered in crude arrangement which isn't attainable for the investigation. Highlight extraction includes diminishing the measure of assets required to portray an extensive arrangement of information. When performing investigation of complex information one of the serious issues originates from the quantity of factors included. Examination with an expansive number of factors by and large requires a lot of memory and calculation control, additionally it might make a characterization calculation overfit to preparing tests and sum up inadequately to new examples.

Highlight extraction is a general term for strategies for building mixes of the factors to get around these issues while as yet portraying the information with adequate exactness. Thus the model of 14 qualities i.e., 14 measurement is diminished to 2 measurement. Records are isolated into two dataset: preparing dataset and testing dataset. Trait "Target" is known as unsurprising property with estimation of "1" for patients with likelihood of having heart assault and estimation of "0" for patients with likelihood of not having heart assault. 66% of information is given for preparing and the staying 33% is given for testing. Model Evaluation is a basic piece of the model advancement process. It finds the best model that speaks to our information and how well the picked model will function later on. Assessing model execution with the information utilized for preparing isn't adequate in information science since it can without much of a stretch produce overoptimistic and overfitted models. The way toward preparing a ML model includes giving a ML calculation (that is, the learning calculation) with preparing information to gain from. The term ML model alludes to the model ancient rarity that is made by the preparation procedure. The model utilized in this structure is ANN (artificial Neural Network). The model is at first fit on a preparation dataset, that is a lot of precedents used to fit the parameters (for example loads of associations between neurons in counterfeit neural systems) of the model. The model is prepared on the preparation dataset utilizing a managed learning strategy. Practically speaking, the preparation dataset

comprises of the objective. The present model is kept running with the preparation dataset and produces an outcome, which is then contrasted and the objective, for each info vector in the preparation dataset. In view of the outcome, the precision will be checked. At last the outcome will be anticipated.

Attributes

Table 1. Input attributes and values data

		Value 0	Value 1	Value 2	Value 3	Value 4	Value 5
1	Age inYear		-	-	-	-	-
2	Sex	Male	Female		-	-	-
1	Age inYear		-	-	-	-	-
2	Sex	Male	Female		-	-	-
3	Chest Pain Type	-	Typical type 1 angina	Typical type 2 angina	Non-angina pain	Asymptomatic	-
4	Fasting Blood Sugar	<120 mg/dl	>120 mg/dl	-	-	-	-
5	Resting	normal	Having ST-T wave abnormality	Showing probable or definite left ventricular hypertrophy	-	-	-
6	Exercise induced angina	No	Yes				
7	Slope – the slope of the		Unslopin	Flat	Down sloping		
8	CA – number of major vessels colored by fluoroscopy		0	1	2	3	-

9	Thal – the heartstatus		Normal	Fixed defect	Reversible defect
10	TrustBloodPressure	Resting blood pressure			
11	Cholesterol	Serum Cholesterol			
12	Halacha	Maximum Heart rate achieved			
13	Oldpeak	ST Depression Induced by e			
14		Heart Disease	Present	No	Yes

4.1 Module 1

The first module consist of the GUI which helps us to take in real time data from the end user being the doctors or the assistant doctors. The input to this module is the 13 attributes which are fed into the graphical user interface. The newly entered data will first be analyzed and predict the condition of the patients heart. We also get the probability of the having an heart disease or not.

4.1.1 Theoretical framework

4.1.1.1 GUI

The graphical UI (GUI) is a type of UI that enables clients to connect with electronic gadgets through graphical symbols and visual pointers, for example, optional documentation, rather than content based UIs, composed order marks or content route. GUIs were acquainted in response with the apparent soak expectation to absorb information of order line interfaces (CLIs), which expect directions to be composed on a PC console.

4.1.1.2 PYQT5

PyQt is a Python authoritative of the cross-stage GUI toolbox Qt, executed as a Python module. PyQt is free programming created by the British firm Riverbank Computing. It is accessible under comparative terms to Qt renditions more established than 4.5; this implies an assortment of licenses including GNU General Public License (GPL) and business permit, yet not the GNU Lesser General Public License (LGPL). PyQt underpins Microsoft Windows just as different kinds of UNIX, including Linux and MacOS (or Darwin). PyQt executes around 440 classes and more than 6,000 capacities and methods including: a considerable arrangement of GUI gadgets classes for getting to SQL databases (ODBC, MySQL, PostgreSQL, Oracle, SQLite) QScintilla, Scintilla-based rich content manager gadget information mindful gadgets that are consequently populated from a database a XML parser SVG support classes for installing ActiveX controls on Windows (just in business form).

4.1.1.3 PROBABILITY

Probability is the measure of the likelihood that an event will occur. Probability quantifies as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur. In probability theory, conditional probability is a measure of the probability of an event (some particular situation occurring) given that another event has occurred. If the event of interest is *A* and the event *B* is known or assumed to have occurred, "the conditional probability of *A* given *B*", or "the probability of *A* under the condition *B*", is usually written as P(*A* | *B*), or sometimes $P_B(A)$ or P(*A* / *B*).

Equations

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

4.2 Module 2

In module 2 we inculcate the following algorithms and analyse which among the algorithms are giving a better accuracy constantly with varied number of datasets.We train our model with the datasets with all the following algorithms which will predict the heart condition of the patients.

4.2.1 Theoretical framework

4.2.1.1 SVM

In the present AI applications, bolster vector machines (SVM) are considered a must attempt—it offers one of themost vigorous and precise techniques among all outstanding calculations. It has a sound hypothetical establishment, requires just twelve models for preparing, what's more, is heartless toward the quantity of measurements. What's more, proficient strategies for preparing SVM are additionally being created at a quick pace. In a two-class learning task, the point of SVM is to locate the best grouping capacity to recognize individuals from the two classes in the preparation information. The measurement for the idea of the "best" grouping capacity can be acknowledged geometrically. For a straightly divisible dataset, a straight characterization work compares to an isolating hyperplane f (x) that goes through the center of the two classes, isolating the two. When this capacity is decided, new information occasion xn can be ordered by basically testing the indication of the capacity f (xn); xn has a place with the positive class if f (xn) > 0.

Since there are numerous such straight hyperplanes, what SVM moreover ensure is that the best such capacity is found by expanding the edge between the two classes. Naturally, the edge is characterized as the measure of room, or partition between the two classes as characterized by the hyperplane. Geometrically, the edge compares to the briefest separation between the nearest information focuses to a point on the hyperplane. Having this geometric definition enables us to investigate how to expand the edge, so that despite the fact that there are a limitless number of hyperplanes, just a couple qualify as the answer for SVM. The motivation behind why SVM demands finding the greatest edge hyperplanes is that it offers the best speculation capacity. It permits not just the best order execution (e.g., exactness) on the preparation information, yet additionally leaves much space for the right order of the future information. To guarantee that the greatest edge hyperplanes are really discovered, a SVM classifier endeavors to boost the accompanying capacity concerning w and b

$$L_P = \frac{1}{2} \parallel \vec{\mathbf{w}} \parallel - \sum_{i=1}^{t} \alpha_i y_i (\vec{\mathbf{w}} \cdot \vec{\mathbf{x}}_\mathbf{i} + b) + \sum_{i=1}^{t} \alpha_i$$

where *t* is the number of training examples, and αi , $i = 1, \ldots, t$, are non-negative numberssuch that the derivatives of *L P* with respect to αi are zero. αi are the Lagrange multipliersand *L P* is called the Lagrangian. In this equation, the vectors **_w** and constant *b* define thehyperplane.

4.2.1.2 NAIVE BAYES

Given a lot of items, every one of which has a place with a known class, and every one of which has a known vector of factors, our point is to develop a standard which will enable us to allot future items to a class, given just the vectors of factors depicting the future articles. Issues of this sort, called issues of regulated grouping, are universal, and numerous techniques for building such standards have been created. One significant one is the gullible Bayes strategy—additionally called numbskull's Bayes, straightforward Bayes, and autonomy Bayes. This strategy is significant for a few reasons. It is exceptionally simple to build, not requiring any entangled iterative parameter estimation plans. This implies it might be promptly connected to enormous information sets. It is anything but difficult to decipher, so clients untalented in classifier innovation can comprehend why it is making the arrangement it makes. Lastly, it regularly does shockingly well: it may not Top 10 calculations in information mining 25 be the most ideal classifier in a specific application, however it can typically be depended on to be vigorous and to do great. General talk of the credulous Bayes technique.

Principle

For accommodation of work here, we will accept only two classes, marked I = 0, 1. Our point is to utilize the underlying arrangement of items with known class enrollments (the preparation set) to develop a score to such an extent that bigger scores are related with class 1 objects (state) and littler scores with class 0 objects. Grouping is then accomplished by contrasting this score and a limit, t. On the off chance that we characterize P(i |x) to be the likelihood that an article with estimation vector x = (x1, . . . , x p) has a place with class I , then any monotonic capacity of P(i |x) would make a appropriate score. Specifically, the proportion P(1|x)/P(0|x) would be reasonable. Rudimentary likelihood discloses to us that we can decay P(i |x) as corresponding to f (x|i)P(i), where f (x|i) is the restrictive dissemination of x for class I items, and P(i) is the likelihood that an article will have a place with class I in the event that we know nothing further about it (the 'earlier' likelihood of class I). This implies the proportion progresses toward becoming

$$\frac{P(1|x)}{P(0|x)} = \frac{f(x|1)P(1)}{f(x|0)P(0)}.$$

4.2.1.3 MLP

A multilayer perceptrons, more formally. A MLP is a finite directed acyclic graph. The nodes that are not arget of any connection are called input neurons. A MLP that should be applied to input patterns of n dimension must have n input neurons, one for each dimension. Input neurons are typically enumerated as neuron1,neuron2,neuron3.The nodes that are no source of any connection are called output neurons. A MLP can have more than one output neuron. The number of output neurons depends on the way the target values(desiredvalues) of thetraining patterns are describedall nodes that are neither input neurons nor output neurons are called hidden neurons.Since the graph is acyclic, all neurons can be organized in layers, with the set of input layers being the first layer.Connections that hop over several layers are called shortcut.Most MLPs have a connection structure with connections from all neurons of one layer to all neurons of the next layer without shortcutsall neurons are enumerated .Succ(i) is the set of all neurons j for which a connection i→j exists.Pred(i) is the set of all neurons j for which a connection j→i existsall connections are weighted with a real number. The weight of the connection i→j is named wji ,all hidden and output neurons have a bias weight. The bias weight of neuron i is named wi0.

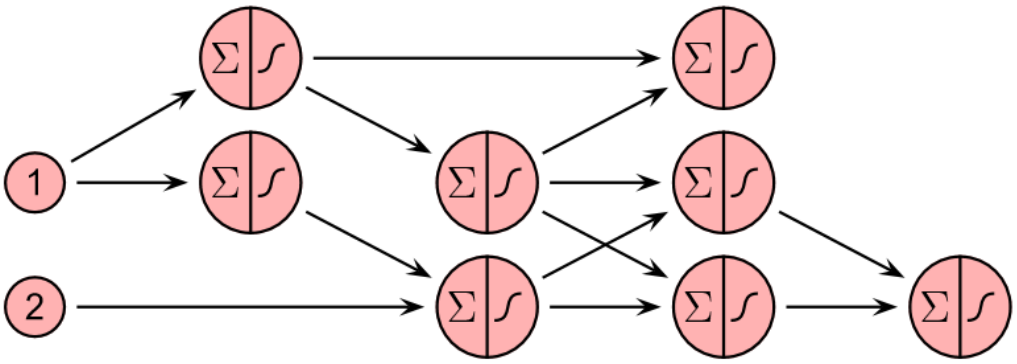

Fig 2:ANN Architecture

- apply pattern ~x=(x1,x2)T .
- calculate activation of input neurons: ai←xi.
- propagate forward the activations: step by step.
- read the network output from both output neurons.

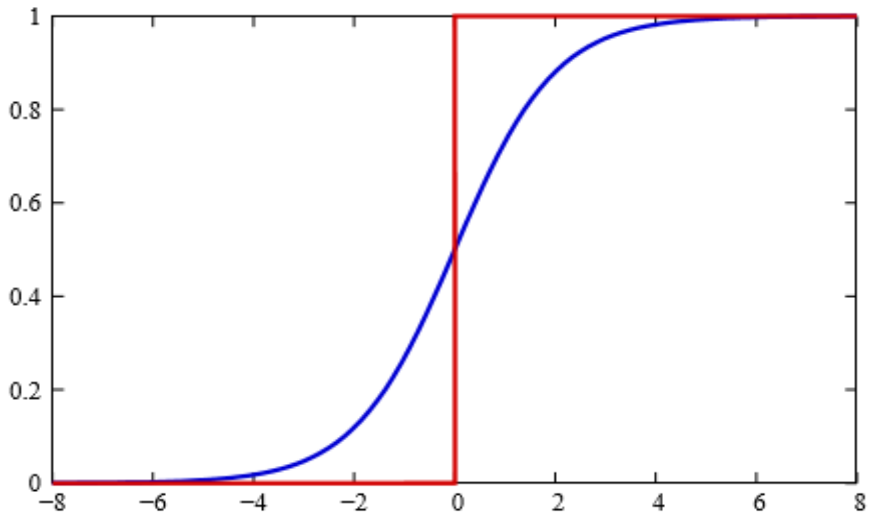

Fig 3: Statistical graph showing the ANN

Require: pattern ~x, MLP, enumeration of all neurons in topological order

Ensure: calculate output of MLP

1: for all input neurons i do

2: set ai←xi

3: end for

4: for all hidden and output neurons i in topological order do

5: set neti←wi0 + Pj ∈ Pred(i) wijaj

6: set ai←flog(neti)

7: end for

8: or all output neurons i do

9: assemble ai in output vector ~y

10: end for

11: return ~y

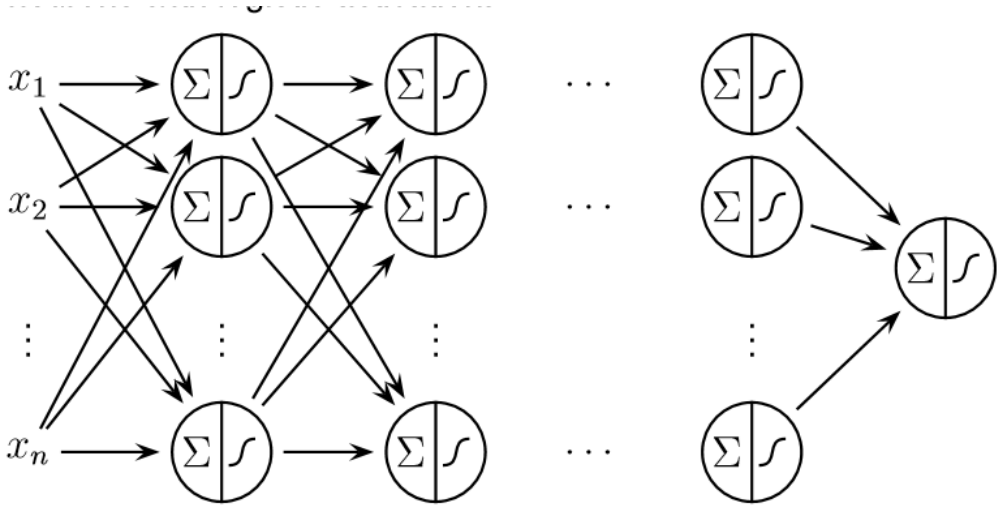

Fig 4: ANN layers

A multi layer perceptrons (MLP) is a finite acyclic graph. The nodes are neurons with logistic activation

Neurons of i-th layer serve as input features for neurons of i+1th layer .Very complex functions can be calculated combining many neurons.

V. CONCLUSION

In implementing the module 1 and module 2 we can give an additional support system to the cardiologist, where in real time doctors tend to forget to check on parameters accounting to an heart disease. This model can help the doctor to predict and prescribe the right medication.As we are training every dynamic data after its prediction we are also getting in huge amount of data which accounts to different,unusual,peculiar,data which came account to the prediction in the future. The later surveys can be taken to find patterns which can be really useful to the research and development field.Prediction of myocardial infarction with the probability of having an heart disease is the outcome of this model.

VI. ACKNOWLEDGMENT

The authors thank their Management and Institute K.S Institute of Technology for providing them the resources and platform for showcasing their idea.They would also like to thank their Prof and HOD, Dr. Rekha B Venkatapur and all lecturers and professors for motivating them into featuring this research.

REFERENCE

[1] Cao, L., Zhang, Z., Gorodetsky, V., & Zhang, C. Interaction between agents and data mining, 2008

[2] Davidson, Ian. “Understanding K-Means Nonhierarchical Clustering”, SUNY Albany-Technical Report 2002.

[3] http://www.ijarcce.com/

[4] Xindong Wu · Vipin Kumar · J. Ross Quinlan “Top 10 algorithms in data mining”, 4 December 2007,Springer-Verlag London Limited 2007.

[5] Igor Kononenko, “Semi-Naïve Bayesian Classifier”,Springer Berlin Heidelberg, March 6–8, 1991,p 206-219

[6] K. Ming Leung, “Naive Bayesian Classifier”, November 28, 2007

[7] Peng, Y., Kou, G. Vol. 7, Issue: 4, Page 639-682, 2008.International Journal of Information Technology and Decision Making System.