
# Data Mining In Health Care Industry

COURSE PROJECT ON REASERACH IN COMPUTER SCIENCE

UNDER THE GUIDELINES OF

Prof. Dr. Lucien Ngalamou

DONE BY

Ektaben Patel

*5009,corning ct,plainfield IL,60586*

*1st semester of master of computer science*

*Lewis university,Romiovilee*

ektabenppatel@lewis.edu

**Abstraction –In this article I want to e discuss about the data mining , how what is data mining? , how data mining is work? , how it is useful in today's life , everywhere it should be use, and main purpose is how the data mining is important for health care industry. The main purpose of data mining to extract the hidden predictive information from the databases which are large. The process will include to divide a dataset into mutually groups in such a way that each group which are**

**"near" and different group which are far from each other and the distance is measured with specific variables that what we are trying to predict. For example, classification deals to divide a database into groups with respect to credit worthiness values like "good" and "bad".**

**And in health care industry data mining is more useful, data mining applications are more beneficial to all the parties who are linked with health care industry. For example in healthcare insurer's data mining will help to detect the frauds and how the health care organizations make the customer relationships management decisions, and what type of best treatment is given to the affected patient with affordable healthcare services.**

**Keyword:**

**Data Mining, health care industry, weka tool, classification, Data set into mutually group , Data Mining on Heart Diesis , Healthcare services, Clustering,K-Mean.**

## i. INTRODUCTION

In 1989 the concept of Knowledge Discovery in Databases it is also known as KDD is introduce by Gregory Piatetsky-Shapiro. after that in 1995 this KDD became the annual ACM SIGKDD Conference on Data Mining . from that to till now data mining is very important for every business and industry level by doing separation on the data from database and making impressive decision.

Data mining means mining the data from huge collection of data. Means that we can do the separation into the huge collection of unused data to get the knowledgeable data.

The ability to data mining medical records has become incredibly important, however records are not as "minable" as data scientist would like to think. As we shown then most of the data is inaccessible.

Good suggestions for mining data, include sequential modeling. We can mine thousands of EHRs that summarizes symptoms and diagnoses and how they correlate with each other. This model could be used to provide patients will mapping of how these parts interact. If we tack the combination of sequential modeling and information integration then it also helpful to knowing about that , how the data interacts are also important components of mining medical data. Information integration encompasses laboratory findings, physiological data, and psychology data. Interactive exploration would mine patients who are not within the researcher's target group. This part of the model would allow people to filter and visualization particular areas of interest.

Data mining medical data is useful for all patients. Why - preventive health care opportunities. Patients and agencies will be able to track treatments, share information between other agencies and track their population.

Data mining, is important because we are looking for tracking clients' Heart disease suppression, meaning how low is the virus in their system.

Like this article points out, an important tool to develop a model that will be able to correlate symptoms, provide better Solution to patients and to see whether analysis efforts are reducing patients' risk of them getting sicker.

## ii. OBJECTIVE

In health care industry the large amount of data generated is in bunch of quantity and it is very complicated to process and analyze by traditional methods. Because of data mining all those data is transformed into useful information for decision making. Data mining and its application helps in within healthcare in such areas like evaluation of treatment effectiveness, and management of healthcare, relationship between about the customer and healthcare units, and the detection of abuse and fraud.

Now a days there is development for a Data mining of a digital infrastructure with the hopes of effectively and efficiently which is provide patients with the best health care. now for any medical center their primary goal is to provide the peasant to good, satisfiable or quality services and avenues to access health care.

Recently, every hospital and medical center are move towards using electronic health records; in order, to better keep track of peasant data.

Data mining is mainly used by the companies who has customer focus, like retail communication, financial organization.
Mainly it deals with the company internal and external relationships internal things like pricing, how to position a particular product and skill of the staff. And the external factors like customer areas, economic indicators and outside competition for the customers. And it will determine the how the sales and customer satisfaction and profits are impacting.

### A. CLASSIFICATION

Actually, it is one of the straightforward method. It is also called as Classification trees or decision trees. And this Algorithm will create a gradually procedure to determine the output of every new data instance. Here a tree of each node in the tree shows spot all the decisions are mainly based on the input provided and will move to the next node until we find a leaf that indicates the predicted output.

Most of the hospital maintains the information like billing system, patient information, information about the staff and some basic simple statistics. They are only able to answer the few questions like approximately how many patients will visit in every year. But they don't have any information about the complex issues. So That's why now a days they are used data mining techniques.

### B. CLUSTERING

It allows a user to make data into groups to determine the patterns of the data. It has a n advantage when the data set is defined and a general pattern needs to determine the data patterns. And we can create a specific number of groups depending on Hospital needs. one benefit of clustering over classification is that each attribute.

### iii. WEKA TOOL

Weka is a machine learning tool which is used for calculations information mining assignments. The calculations can either be connected specifically to a dataset or called from your own particular Java code. Weka contains devices for information pre-preparing, grouping, relapse, bunching, affiliation principles, and perception. It is likewise appropriate for growing new machine learning plans.

Discovered just on the islands of New Zealand, the Weka is a flightless fledgling with a curious nature. The name is articulated this way, and the winged animal sounds this way.

Weka is open source programming issued under the GNU General Public License. It Clarifies how machine learning calculations for information mining work.

Helps you look at and assess the consequences of various procedures. Covers execution change procedures, including input preprocessing and consolidating yield from various strategies.
Highlights top to bottom data on probabilistic models and profound learning.

Gives a prologue to the Weka machine learning workbench and connections to calculation executions in the product.
Description of the Dataset: This catalog contains 4 databases concerning coronary illness finding.

All properties are numeric-esteemed. The information was gathered from the four after areas:

1. Cleveland Clinic Foundation (Cleveland. Data)

2. Hungarian Institute of Cardiology, Budapest (Hungarian. Data)

3. V.A. Therapeutic Center, Long Beach, CA (long-shoreline va.Data)

4. College Hospital, Zurich, Switzerland (Switzerland. Data)

Every database has a similar example arrange. While the databases have 76 crude characteristics, just 14 of them are really utilized. In this way, I've taken the freedom of making 2 duplicates of every database: one with every one of the characteristics furthermore, 1 with the 14 traits really utilized as a part of past investigations.
The creators of the databases have asked: that any productions coming about because of the utilization of the information incorporate the names of the chief agent in charge of the information gathering at every organization. They would be:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.



This database contains 76 Attributes, however all distributed tests allude to utilizing a subset of 14 of them. Specifically, the Cleveland database is the special case that has been utilized by ML scientists to this date. The "objective" field alludes to the nearness of coronary illness in the patient. It is number esteemed from 0 (no nearness) to 4. Tries different things with the Cleveland database have focused on basically endeavoring to recognize nearness (values 1,2,3,4) from nonattendance (esteem 0).

The names and government disability quantities of the patients were as of late expelled from the database, supplanted with sham qualities.

One record has been "prepared", that one containing the Cleveland database. Every one of the four natural records likewise exist in this index.

To see Test Costs (gave by Peter Turney), please observe the envelope "Expenses"

### iv. ATTRIBUTE INFORMATION:

There are so many various type of attribute are effected to Heart Disease problem but I want to discuss some of them attribute which are more effected to heart Diseas problem , and how to solve this problem.

### A. KEY ATTRIBUTES USED

1 age: age in years

2 sex: sex (1 = male; 0 = female)

3 cp: chest pain type
Value 1: typical angina

Value 2: atypical angina
Value 3: non-anginal pain

Value 4: asymptomatic

4 trestbps: resting blood pressure (in mm Hg on admission to the hospital)

5 chol: serum cholestoral in mg/dl

6 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7 restecg: resting electrocardiographic results

Value 0: normal

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.

8 thalach: maximum heart rate achieved

9 exang: exercise induced angina (1 = yes; 0 = no)

10 old peak = ST depression induced by exercise relative to rest

11 slope: the slope of the peak exercise ST segment IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009 233

Value 1: up sloping

Value 2: flat

Value 3: down sloping

12 ca: number of major vessels (0-3) colored by fluoroscopy

13 thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

14 num: diagnosis of heart disease (angiographic disease status)

Value 0: < 50% diameter narrowing

Value 1: > 50% diameter narrowing

### v. DATA SET DESCRIPTION:

This is the resource from where I took the data and doing analysis.


Decision tree Algorithms: it is one of the method used to construct a model of decisions made based on the actual values of attributes in the data. This decision trees are trained on data for classification and regression problems. Is it one of the big favorite in machine learning and is also often fast and accurate.

### A. THE MOST POPULAR DECISION TREE ALGORITHMS ARE:

Classification and Regression Tree (CART)

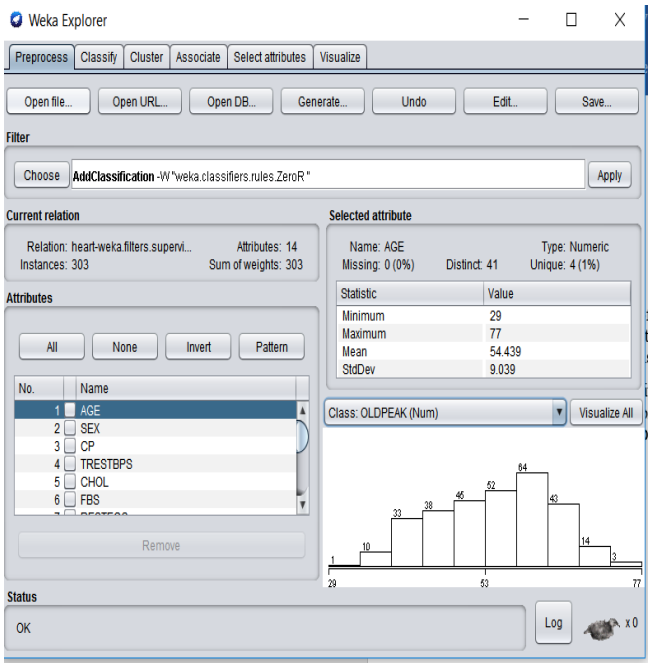
Iterative Dichotomiser 3 (ID3)

C

Decision Stump

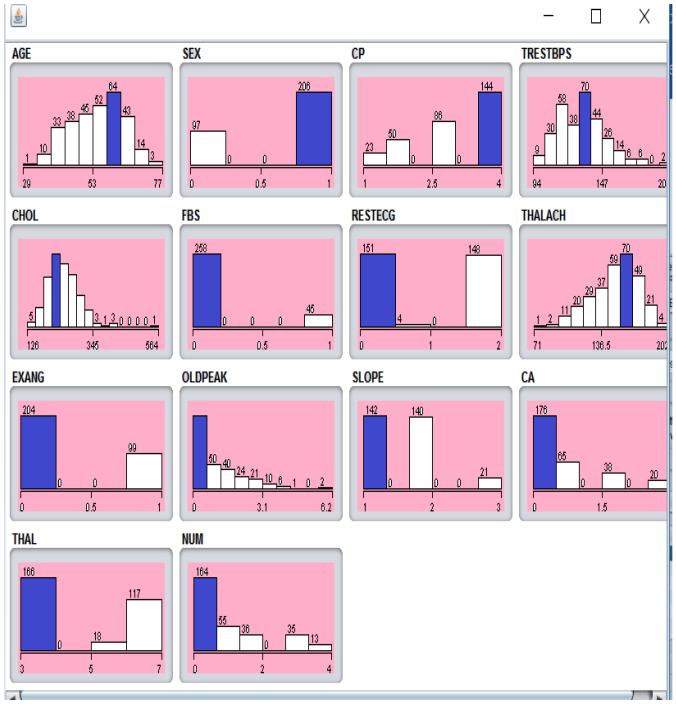
M5

### B. CONDITIONAL DECISION TREES




Now we are analyzing each and every attribute to analysis what are the actual reasons for the heart disease.
The reasons may be multiple it may be age, heart rate level , may be the person is diabetic.

### C. We can see all the reports in the below




## vi. Result:

After analyzing the data we understand that who are affecting with Heart diseases:

- Mostly the age group above 52 are affecting that to males are affecting.
- Most of the people are affecting with the chest pain type is Asymptomatic(I.e unfamiliar symptoms to the patient regarding the pain)
- The persons who is having more cholesterol will effect the heart disease.
- If fasting blood sugar is high that is also the reason for heart disease.
- By looking at the ECG results will know the severity of the heart disease.
- The people who are doing exercise are less victims to the heart disease.

## vii. Data mining model

Data mining model means mining the model which is created by passing an algorithm to data and this is a group or bunch of data and patterns , that can be applied to new data for produce the forecast and make a conclusion reached on the basis of evidence and reasoning about relationships.

There is a four type of model in data mining which are Classifying, Regression, Clustering, Decision tree.

### A. Classification

Actually, it is one of the straightforward method. It is also called as Classification trees. And this Algorithm will create in succession order to procedure to determine the output of every new data instance.

In short classification means structure the data from the unstructured behavior. In industry and in business or we can say that every where there is huge amount of data is available and in bad format so data mining is help full to classify that data into the stretchered or a use full format. So it can be used for decision making point of view.

### B. Regression

Regression is one of the data mining model or function which estimate that what will be happen in the future in our analysis it is predict that AGE, CP, SEX, fbs ,restecg , thalach ,exang Etc,

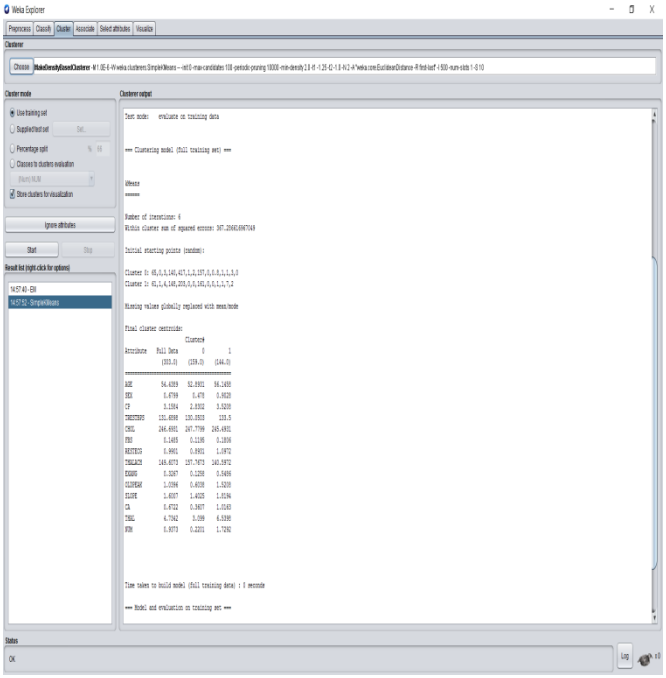
Regression model is also sub divided into three model which is linear, polynomial, and logistic regression. In linear and polynomial regression having a continuous spectrum of values. For this Bothe type of method we can not use the normalization Method.

## C. CLUSTERING

Usually will make data in groups in Clustering and it will determine the patterns of the information. It has an advantage when the data set is defined and a general pattern needs to determine the data patterns. And we can create a specific number of groups depending on business needs. One benefit of clustering over classification is that each attribute.

This is one of the clustering technique which is used to work with data.

K means is one of the simple algorithm which solve the clustering problem. It will divide the data into number of clusters. And these points act as centriods. All the data is grouped to the particular centriods. And we need to repeat the same process to get the result.



## C. DECISION TREE

decision tree is basically the combination of classification and regression and they give the result in the form of tree. It divide the data set into subset and accede an associated decision tree is progressively developed.

And final result should be with two or more branches it is also called decision node or it should be facing the classification it also called leaf node.

### viii. RUN INFORMATION

I want to be talk about the running information which we need when we run this clustering with k-means algorithm on the weka tool , to get particular result.

.

### A. SCHEME

weka.clusterers. Simple K-Means -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: heart
Instances: 303
Attributes: 14
AGE
SEX
CP
TRESTBPS
CHOL
FBS
RESTECG
THALACH
EXANG
OLDPEAK
SLOPE
CA
THAL
NUM

Test mode: evaluate on training data

### ix. CLUSTERING MODEL (FULL TRAINING SET)

### A. K MEANS

Number of iterations: 6

Within cluster sum of squared errors: 367.286616967049

Initial starting points (random):

Cluster 0: 65,0,3,140,417,1,2,157,0,0.8,1,1,3,0

Cluster 1: 61,1,4,148,203,0,0,161,0,0,1,1,7,2

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0 seconds

### B. MODEL AND EVALUATION ON TRAINING SET

Clustered Instances:

0 159 ( 52%)

1 144 ( 48%)

### C. Final Centro

| Attribute | Full Data | 0 | 1 |
|---|---|---|---|
| | (303.0) | (159.0) | (144.0) |
| | | | |
| AGE | 54.4389 | 52.8931 | 56.1458 |
| SEX | 0.6799 | 0.478 | 0.9028 |
| CP | 3.1584 | 2.8302 | 3.5208 |
| TRESTBPS | 131.6898 | 130.0503 | 133.5 |
| CHOL | 246.6931 | 247.7799 | 245.4931 |
| FBS | 0.1485 | 0.1195 | 0.1806 |
| RESTECG | 0.9901 | 0.8931 | 1.0972. |
| THALACH | 149.6073 | 157.7673 | 140.5972 |
| EXANG | 0.3267 | 0.1258 | 0.5486 |
| OLDPEAK | 1.0396 | 0.6038 | 1.5208 |
| SLOPE | 1.6007 | 1.4025 | 1.8194 |
| CA | 0.6722 | 0.3607 | 1.0163 |
| THAL | 4.7342 | 3.099 | 6.5398 |
| NUM | 0.9373 | 0.2201 | 1.7292 |

## x. Conclusion.

By using the dataset the predictions for heart Disease are extracted. In this we got two clusters that is cluster 0 and cluster 1 And looking at the below table cluster 0 is having most of the data i.e almost 159 instances and cluster 1 contains 144 instances. These are the predictions that can develop the heart attack .

This prediction helps doctors to diagnosis diseases correctly With the attributes. In future, It will also helpful to diagnose the Other technique different diseases like–Cancer type, HIV prediction etc. to know the better prediction results.

Taking decision in Heart disease prediction which is developed by K-means technique. In this model it extracts the hidden information from the heart disease database.

Actually this model is effective to predict the patients who is Suffering with heart disease. Data mining techniques were able to answer few complex questions in any field which is related to data.

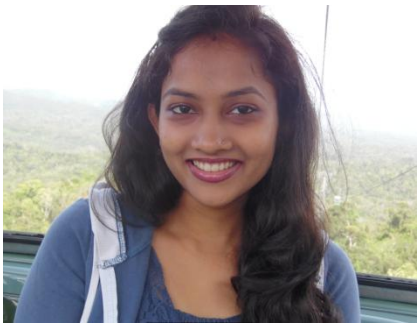
## xi. References


[1] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

[2] https://www.healthcatalyst.com/wp-content/uploads/2014/06/What-is-data-mining-in-healthcare.pdf

[3] http://scikit- learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[4] https://www.youtube.com/watch?v=m7kpIBGEdkI


Ekta Patel was born in Surat , India in 1992. I completed my master in computer application in India from Gujarat Technological University and now i am majoring in computer science in Lewis university, Romeoville.