Heart Diseases Prediction Using WEKA

Asst. lecturer Eng. Tamara Saad Mohamed

Baghdad college of economic sciences university

Eng. Mohammed Hussein Ali

الخلاصة :

نتيجة لنمط الحياة المتسارع اصبحت أمراض القلب تزداد يوما بعد يوم وتجعل حياة الناس في خطر. فاصبحت أمراض القلب من أكثر الأمراض شيوعًا هذه الأيام. لذلك أصبح من الضروري للغاية البحث ومحاولة إيجاد أبسط وأفضل طريقة للتنبؤ بالأمراض مقدمًا لان ذلك قد يساعد حياة الناس على النجاة.

نحاول في هذا البحث الاعتماد على فحص طبي معين يمكن أن يساعد في تقييم أمراض القلب بلاعتماد على بعض الاعراض مثل ضغط الدم وألم الصدر والكوليسترول والسكر في الدم إلى العمر والجنس للعثور على مريض مصاب وما هي أفضل خوارزمية لاستخدامها من أجل الحصول على دقة جيدة.

باستخدام تقنيات ال data mining وبعض الخوارزميات المتوفرة في ال WEKA 3.8.1(Waikato Environment for Knowledge analysis) سنقوم بعمل مقارنه و تحليل بين النتائج المستحصلة باستخدام نوعين من الخوارزميات لتحديد اي الخوارزميتان افضل واكثر دقه للتنبؤ بامراض القلب.

Abstract :

Consequent to the life style and day by day heart diseases increasing and make people’s life at risk . Heart diseases becomes one of the most common diseases these days . Though it become very necessary to search and try to find the simplest and best way to predicate diseases in advance though that could help survive the life of people .

This paper try to depend on a particular medical examination that could help predicate the heart diseases such as blood pressure , chest pain , cholesterol , blood sugar along to age and sex to find weather patient has disease or not and what is the best algorithm to use in order to get a good accuracy .

By using data mining technique and algorithm that available in WEKA 3.8.1(Waikato Environment for Knowledge Analysis) tool we are going to exam two algorithms; the decision tree J84 and Naïve Bayesian , then analyze and compare the result of both to find the most accurate one . The prediction of heart diseases survivability has many challenging issues in several related fields.

1. Introduction :

Data mining is a method that helps extracting knowledge from a huge data by analyzes data with number of algorithms that could help extract the useful data from big data . Data mining gives a lots of algorithms to groups the data and make the data more useful ,to achieve this goal by depending on specific technique to examine and analyze the data to collect the better knowledge . In health care data mining become one of the most useful technique , the industry of health care generates large amount of data about patients ,diagnosis , diseases , and so on . In a health care field the quality of services is a common issues which the medical institutes suffered from .the quality involves the diagnoses the diseases correctly and providing the right advises , the inferior diagnoses ends with fetal results which faced the medical institutes .

2. problem description :

The data set which used in this paper is for diseases of heart , the data set which available in heart diseases is ARFF(Attribute-Relation File Format) collected from UCI(Unique Client Identifier) repository . The dataset describes the risk factor of heart diseases . The attribute represent the binary class attribute

:class <50 means there is no disease ;class >50 indicates increase the possibility of the heart diseases .

The goal of this paper is to decrease the possibility of getting risk of heart diseases by predict the heart disease in advance by using classification in data mining technique from other attributes in dataset .

The software which proposed to be used to implement this work will be WEKA 3.8.1 .

3. literature overview :

Various investigations have been done , that have center around analysis of heart maladies. The importance and the critical issues of the heart diseases on people life , this topic was always the focus of interest of researchers to get the best technique that could help predict the heart diseases . They have utilized and connected diverse data mining systems for finding and accomplished distinctive probabilities for various strategies and algorithms [4][5] .

A. NITI et .al. : focused on using of neural network to get prediction of the heart diseases by depending the percent of sugar and blood pressure with some other attributes . The dataset contains record with 13 attributes in each record . The regular networks for example the neural network which woks with the back propagation method is used for examine and training the data [5].

B. The inconvenience of relating the constrained associated rules for heart diseases expectation was contemplated via Carlos Ordonez . the resultant dataset contains records of patients having heart diseases .There are three imperatives were acquainted with abatement the quantity of habits . they are as per the following

- **a)** The attributes have to appear on only one side of the rule .
- **b)** Separate the attributes into groups , i.e. uninteresting groups .
- **c)** In a rule , there should be limited number of attributes .
    - **a)** The attributes need to show up on just a single side of the rule .
    - **b)** Separate the attributes into gatherings , for example uninteresting gatherings .
    - **c)** There ought to be set number of attributes in each rule .

The result of this method is two groups of rules , the presence or absence of heart diseases [4].

C. Latheparathiban et .al. : focused on using basic of coactive Neuro-Fuzzy Inference System (CANFIS) for prediction of heart diseases ; the CANFIS demonstrate utilizes neural system abilities with the genetic algorithm and fuzzy logic[5] .

D. Kiyong Noh et .al. : focused on using an extraction of Muli-Parametric for classification method advantages by evaluation HRV (Heart Rate Variability) from ECG (Electro Cardio Gram) .the dataset had been used composed of 670 persons , they are classified into two groups , they are patients with heart diseases and normal people , which were used to complete the examination for the affiliated classifier.

E. Akhiljabbar et.al. : focused on using of proficient associative classification algorithm utilizing genetic algorithm for heart diseases prediction . The primary inspiration for utilizing genetic algorithm in the revelation of abnormal state prediction rules is that the discovered rules are highly expectation precision and high intriguing quality of those quantities.

F. ShrutiRantnakar et .al. : focused on using of the genetic algorithm to deduct the set of attributes of (Naïve Bayes) which creates the relations among the attributes .

4. Implementation tools and algorithms :

4.1 classification :

Classification is a method using for arranging the information in the shape of dependent class dependent on certain feature or comparability. This procedure needs extraction and picking of form that is a good portrays to a specific class. Classification is additionally called directed learning, Each case in the dataset appeared to by set of forms or qualities which might be unmitigated or persistent. classification is the way toward create the prototype from the preparation set. The subsequent prototype then is used to anticipate the class name of the examine case[9].

4.2 Algorithms :

4.2.1 Decision tree J48 :

This segment clarifies the classification algorithm J48 and Naïve Bayes. (A.J48 decision tree classifier): J48 is the decision tree based algorithm and it is the augmentation of C4.5. With this procedure a tree is developed to show the grouping procedure in decision tree the interior nodes of the tree indicates a test on a attributes , branch speak to the result of the test, leaf node holds a class mark and the highest node is the root node . The result is produced by decision tree predicts new cases of data[11].

Algorithm J48:

```
Input D // Training data
Output T // Decision tree
DTBUILD (*D) (T – Null) ;
T = Create root node and label with splitting attribute ;
T = Add arc to root node for each split predict and label ;
For each arc do
D = Database created by applying splitting predict to D ;
If stopping point reached for this path , then
T’ = Create leaf node and label with appropriate class ;
Else T’ = DTBUILD (D) ;
T = Add T’ to arc ;
```

While building tree J48 ignores the missing value .J48 allows classification via either decision tree or rules generated from them.

4.2.2. Naïve Bayesian classifier :

Bayesian specification act to monitor learning technique just as measurable strategy for grouping or classification. It is straightforward probabilistic classifier dependent on Bayesian hypothesis with robust unrestraint hypothesis. It is especially fit when the dimensionality of input is high. They can anticipate the probability that a given tuple has a place with a specific class. This classification is named after Thomas Bayes (1702-1761) who proposed the bayes hypothesis. Bayesian equation can be composed as P(H | E) = [P(E | H) * P(H)]/P(E) The essential thought of Bayes' rule is that the result of a speculation or an event (H) can be anticipated dependent on certain confirmations (E) that can be seen from the Bayes' rule[1][3].

4.3 WEKA Tool :

Weka (Waikato environment for knowledge analysis) is a mainstream suite of machine learning programming written in Java, created at the University of Waikato, New Zealand. The Weka suite contains a gathering of representation algorithms and apparatus for data analusis. In Weka dataset ought to be designed to the ARFF position. The Weka Explorer will utilize these naturally in the event that it doesn't perceive a given document as an ARFF record. The pre-process board has offices for bringing in data from database and for pre-handling this information utilizing filtering algorithms. These channels can be utilized to change the data[6][8].

5. Performance examination and results:

5.1 Data understanding :

The fundamental advance in moving toward the issue is to get to know the data. The dataset we are going to test in this task incorporate (14) properties gathered from UCI data archive which are the most impact on heart disease predication, the attributes are :

Attribute information :

- Age
- Sex
- Chest pain type (4 values)
- Resting blood pressure
- Serum cholesterol in mg/dl
- Fasting blood sugar > 120 mg/dl
- Resting electrocardiographic results (values 0,1,2)
- Maximum heart rate achieved
- Exercise induced angina
- Oldpeak = ST depression induced by exercise relative to rest
- The slope of the peak exercise ST segment
- Number of major vessels (0-3) colored by fluoroscopy
- Thal :3 = normal ; 6= fixed defect ; 7= reversible defect

The above attributes are different data types as showing below :
Attributes types :

- Real : 1,4,5,8,10,12
- Ordered :11
- Binary : 2,6,9
- Nominal : 7,3,13

Mining the values of all attributes will help predicate the heart disease in advance and classify weather disease is present or not.

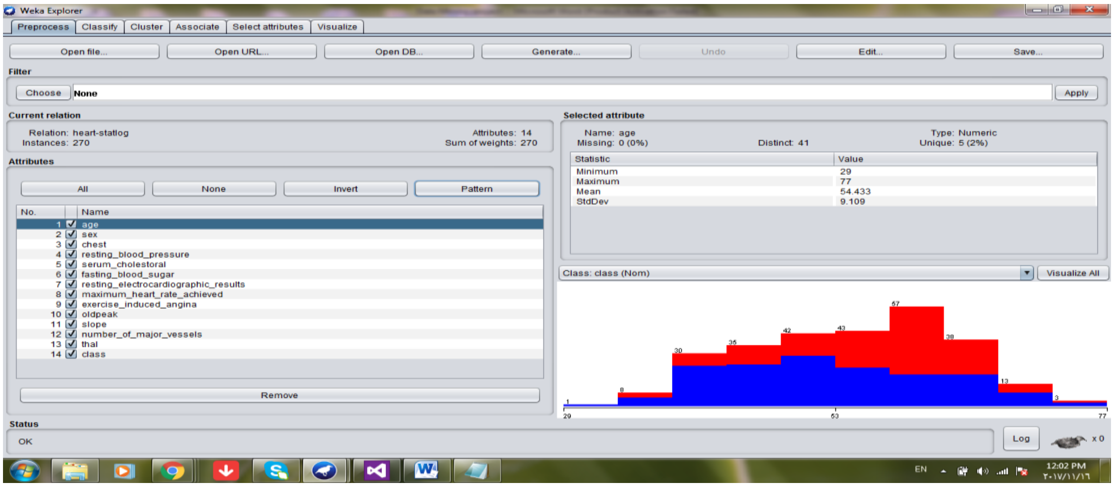
In next section we are going to implement two data mining algorithm to test which one goanna give the accurate result[7].

5.2 Data preprocessing :

Next step in order to mining data is preprocessing data such that transform it into most suitable form to perform the mining algorithm.

- Attribute selection .
- Handling a missing data.
- Eliminate the outlier .

14 attributes selected to preprocessing data with (270) instances and (0) missing values in our data .As shown in fig(1)



Filter

None

Current relation

Relation: heart-statlog
Instances: 270
Attributes: 14
Sum of weights: 270

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 41
Type: Numeric
Unique: 5 (2%)

Statistic	Value
Minimum	29
Maximum	77
Mean	54.433
StdDev	9.109

Class: class (Nom)

Attributes

No.		Name
1	☑	age
2	☑	sex
3	☑	chest
4	☑	resting_blood_pressure
5	☑	serum_cholestoral
6	☑	fasting_blood_sugar
7	☑	resting_electrocardiographic_results
8	☑	maximum_heart_rate_achieved
9	☑	exercise_induced_angina
10	☑	oldpeak
11	☑	slope
12	☑	number_of_major_vessels
13	☑	thal
14	☑	class

Status

OK

Figure (1) Heart diseases dataset preprocessing

The visualization of all selected attributes shown in fig(2)

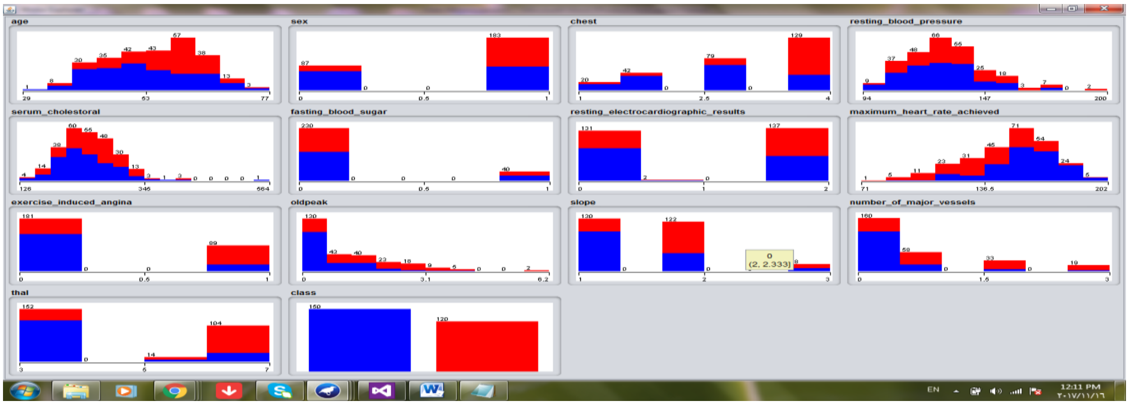


Figure (2) Visualization of all attributes

5.3 Mining data :

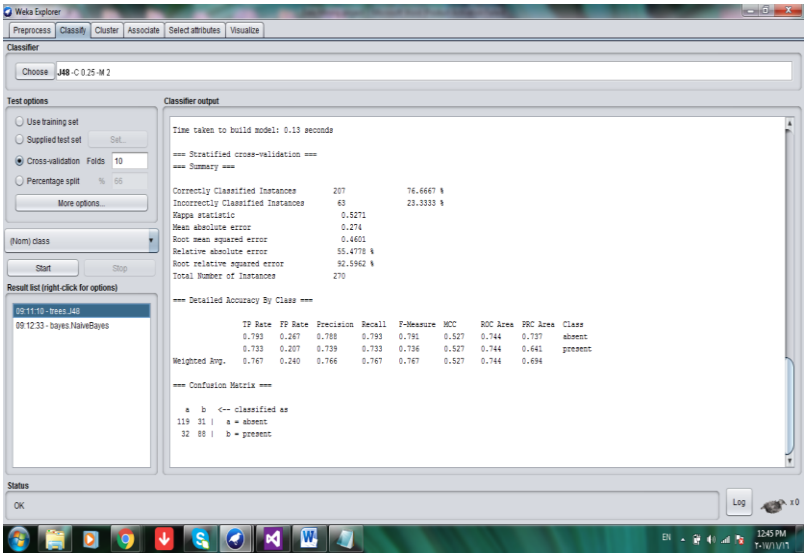
The next step is to mining data with using two classification algorithms and finds the difference in result of both. Exercise performed with WEKA using 10 fold cross validation. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier. The first step is to find the number of instances of heart disease dataset using both J48 and Naïve Bayes classification algorithm.Then calculates the classification accuracy and cost analysis using Confusion Matrix which contains information about actual and predicted classification[4][5].

5.4 Standard terms defined for this matrix as the following:

- True positive –if the result of forecast is p and the real esteem is additionally p than it is called true positive(TP).
- False positive-if agenuin value is n than it is false positive(FP)
- Precision – accuracy is proportion of precision and quality Precision = tp/(tp + fp)
- Recall- proportion of fulfillment and the amount Recall = tp / (tp + fn)

6. Using J48 algorithm :

J48 is a technique for creating a pruned or unpruned C4.5 decision tree. Figure (3) shows J48 tree for heart disease data.



Classifier: J48 -C 0.25 -M 2

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

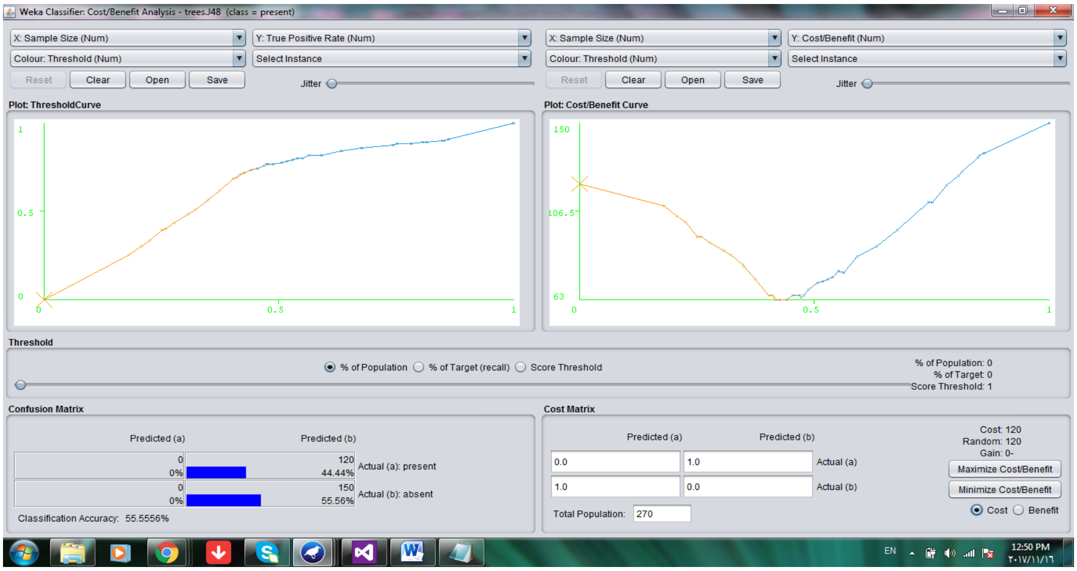
Result list (right-click for options):
09:11:10 - trees.J48
09:12:33 - bayes.NaiveBayes

Status: OK

```
Time taken to build model: 0.13 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               76.6667 %
Incorrectly Classified Instances        63               23.3333 %
Kappa statistic                          0.5271
Mean absolute error                      0.274
Root mean squared error                  0.4601
Relative absolute error                 55.4778 %
Root relative squared error             92.5962 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.793    0.267    0.788      0.793    0.791      0.527    0.744     0.737     absent
                 0.733    0.207    0.739      0.733    0.736      0.527    0.744     0.641     present
Weighted Avg.    0.767    0.240    0.766      0.767    0.767      0.527    0.744     0.694

=== Confusion Matrix ===

   a   b   <-- classified as
 119  31 |   a = absent
  32  88 |   b = present
```

Figure (3) result of J48 tree

The cost analysis for present class of J48 classification algorithm shown in fig (4).The cost analysis for absent class of J48 classification algorithm shown in fig (5) , both types of costs calculated by using WEKA for the heart diseases dataset which includes the 14 attributes selected to processing data which are shown in fig (1) .



Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = present)

X: Sample Size (Num) Y: True Positive Rate (Num) Colour: Threshold (Num) Select Instance

X: Sample Size (Num) Y: Cost/Benefit (Num) Colour: Threshold (Num) Select Instance

Threshold: ● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	120 44.44%	Actual (a): present
0 0%	150 55.56%	Actual (b): absent

Classification Accuracy: 55.5556%

Cost Matrix

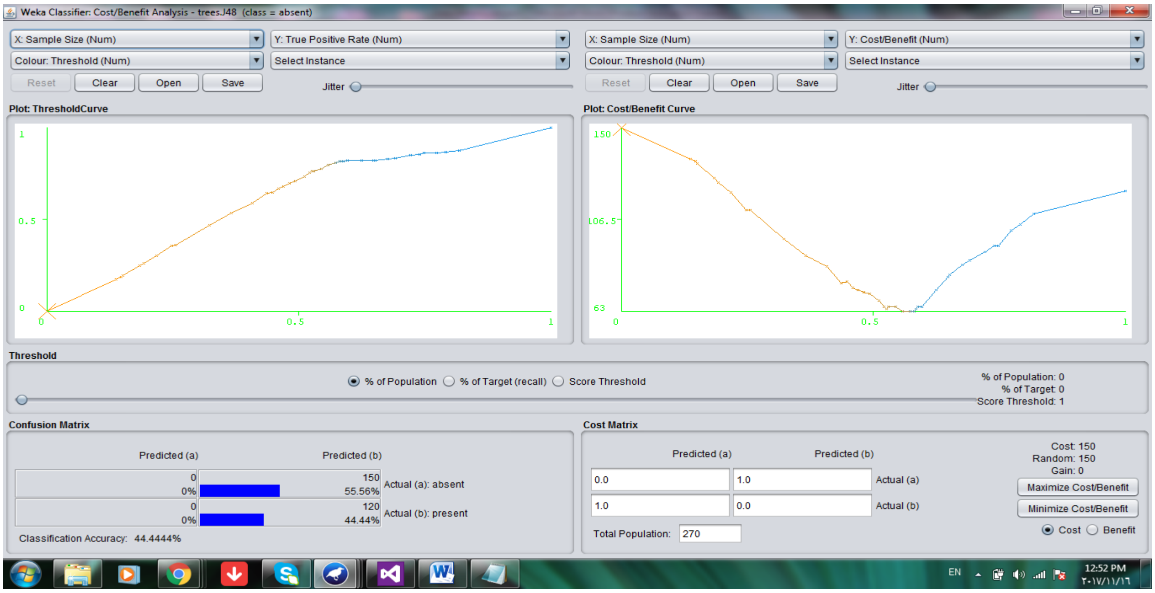
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 270

Cost: 120
Random: 120
Gain: 0-

● Cost ○ Benefit

Figure (4) cost analysis for present class



Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = absent)

X: Sample Size (Num) Y: True Positive Rate (Num) Colour: Threshold (Num) Select Instance

X: Sample Size (Num) Y: Cost/Benefit (Num) Colour: Threshold (Num) Select Instance

Threshold: ● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	150 55.56%	Actual (a): absent
0 0%	120 44.44%	Actual (b): present

Classification Accuracy: 44.4444%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

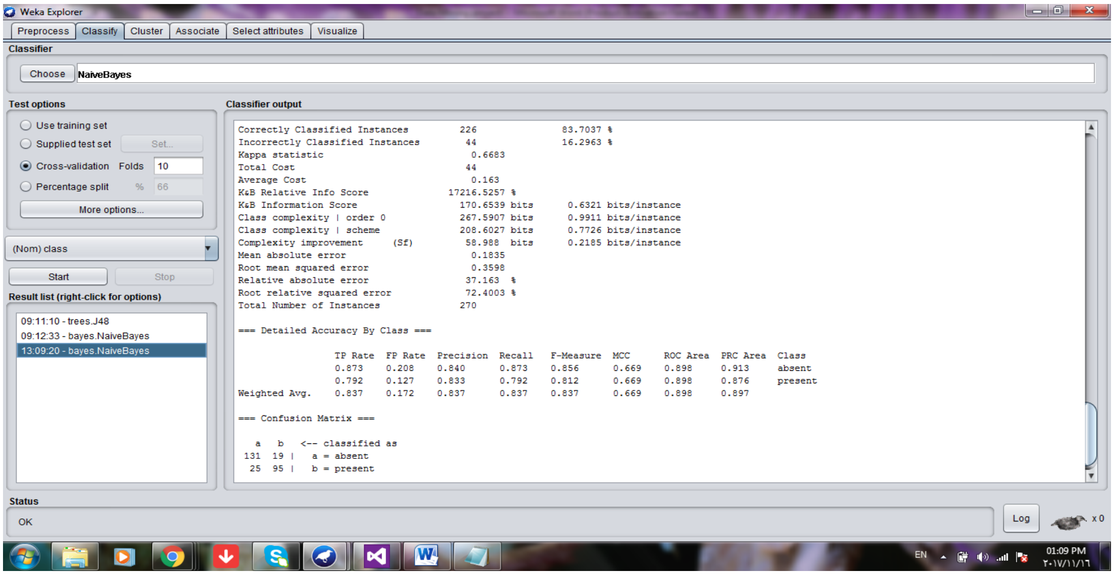
Total Population: 270

Cost: 150
Random: 150
Gain: 0

● Cost ○ Benefit

Figure (5) cost analysis for absent class

- ✓ Result of using Naïve Bayes algorithm :
  Fig (6) shows the applying of Naïve Bayes algorithm and its classification output , fig (7) shows the cost analysis of present class of Naïve Bayes algorithm and fig(8) shown the cost analysis of absent class of Naïve Bayes algorithm.



Classifier: NaiveBayes

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Result list (right-click for options)

- 09:11:10 - trees.J48
- 09:12:33 - bayes.NaiveBayes
- 13:09:20 - bayes.NaiveBayes

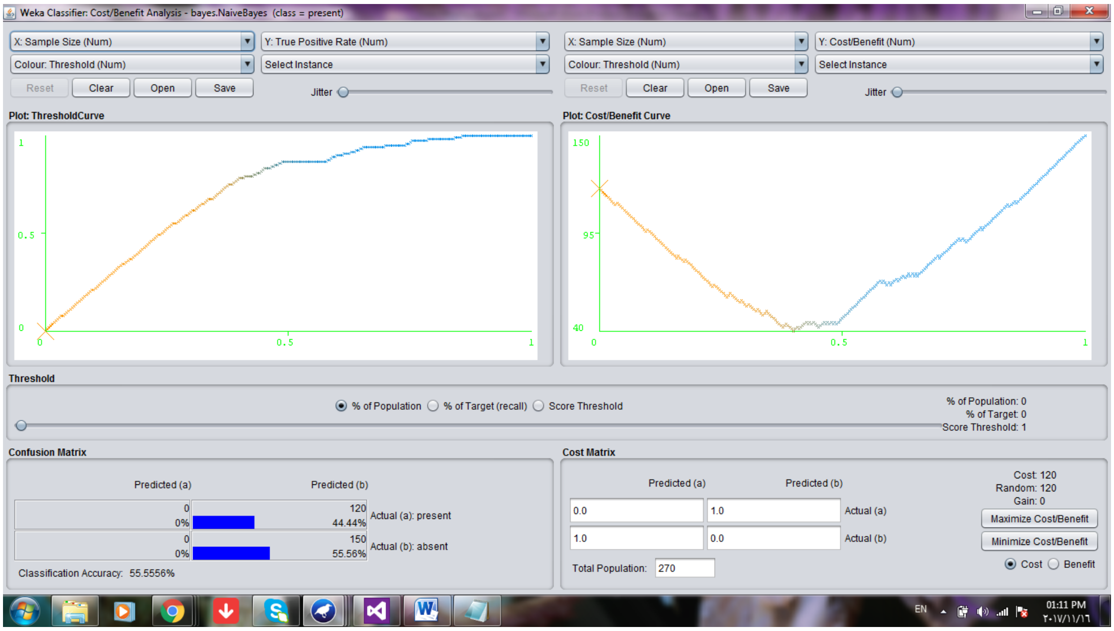
Classifier output

```
Correctly Classified Instances         226               83.7037 %
Incorrectly Classified Instances        44               16.2963 %
Kappa statistic                          0.6683
Total Cost                              44
Average Cost                             0.163
K&B Relative Info Score              17216.5257 %
K&B Information Score                  170.6539 bits      0.6321 bits/instance
Class complexity | order 0             267.5907 bits      0.9911 bits/instance
Class complexity | scheme              208.6027 bits      0.7726 bits/instance
Complexity improvement     (Sf)         58.988  bits      0.2185 bits/instance
Mean absolute error                      0.1835
Root mean squared error                  0.3598
Relative absolute error                 37.163  %
Root relative squared error             72.4003 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.873    0.208    0.840      0.873    0.856      0.669    0.898     0.913     absent
                 0.792    0.127    0.833      0.792    0.812      0.669    0.898     0.876     present
Weighted Avg.    0.837    0.172    0.837      0.837    0.837      0.669    0.898     0.897

=== Confusion Matrix ===

   a   b   <-- classified as
 131  19 |   a = absent
  25  95 |   b = present
```

Status

OK

Figure (6) Naïve Bayes algorithm



X: Sample Size (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance

X: Sample Size (Num)
Y: Cost/Benefit (Num)
Colour: Threshold (Num)
Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	120 44.44%	Actual (a): present
0 0%	150 55.56%	Actual (b): absent

Classification Accuracy: 55.5556%

Cost Matrix

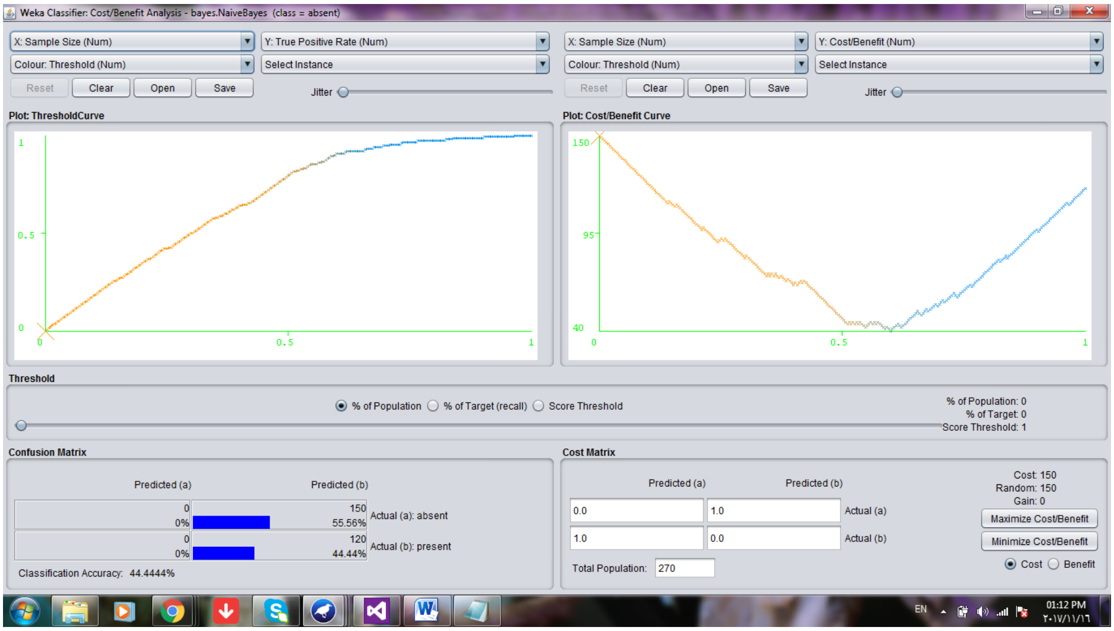
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 270

Cost: 120
Random: 120
Gain: 0

● Cost ○ Benefit

Figure (7) cost analysis for present class



Weka Classifier: Cost/Benefit Analysis - bayes.NaiveBayes (class = absent)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Threshold (Num) | Select Instance

Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	150 55.56%	Actual (a): absent
0 0%	120 44.44%	Actual (b): present

Classification Accuracy: 44.4444%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 270

Cost: 150
Random: 150
Gain: 0

Figure (8) cost analysis for absent class

7. Conclusion :

After applying both of algorithms naïve Bayes and J48 on Heart disease dataset the result are given below:

Evaluation Criteria	Naïve Bayes	J48
Time to achieve model (in sec)	0.02	0.13
Correctly classified instances	226	207
Incorrectly classified instances	44	63
Prediction accuracy	83.70 %	76.66 %

Table (5.1) result comparison of using naïve Bayes and J48

From results given in table(5.1) we notice that time to build the model in Naïve Bayes is less than the time required to build the model in J48, furthermore, accurately characterized occasions are more when utilizing Naive Bayes too forecast exactness is additionally more prominent in Naive Bayes than of J48. Henceforth it is presumed that Naïve Bayes perform superior to of J48 on heart diseases dataset .We can conclude the using of Naïve bayes algorithm is more efficient than J48 according to the selected criteria in table (5.1).

References:

1. A.Goyal, R.Mehta," Performance comparison of Naïve Bayes and J48 classification algorithms" .

2. Efraim Turban, Linda Volonino, Information Technology for Management: Wiley Publication, 8th Edition 2009.

3. Sunita Joshi, Bhuwaneshwari Pandey,Nitin Joshi, Comparative analysis of Naive Bayes and J48 Classification Algorithms published in International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X.(2015).

4. lan H. Witten, Eibe Frank, "Data Mining –Practical Machine Learning Tools and Techniques,"2nd Edition, Elsevier, 2005.

5. Analysis of Classification Algorithms for Heart Disease Prediction and its Accuracies R. Jothikumar and R.V. Sivabalan 1 2 Noorul Islam University, Kumaracoil, Tuckalay, 1 Kanyakumari Dt-629180, Tamil Nadu, India Department of Master of Computer Application, Noorul Islam University, 2 Kumaracoil, Tuckalay, Kanyakumari Dt-629180, Tamil Nadu, India

6. Analysis of Heart Disease using in Data Mining Tools Orange and Weka , Global Journal of Computer Science and Technology: C, Software & Data Engineering, Volume 18 Issue 1 Version 1.0 Year 2018, Online ISSN: 0975-4172 & Print ISSN: 0975-4350

7. Heart Disease Prediction System using Data Mining Method, Keerthana T K #1 #PG student, Dept. of Computer Science, Jyothi Engineering College Thrissur, Kerala, India, International Journal of Engineering Trends and Technology (IJETT) – Volume 47 Number 6 May 2017.

8. DENGUE DISEASE PREDICTION USING WEKA DATA MINING TOOL KASHISH ARA SHAKIL, SHADMA ANIS AND MANSAF ALAM Department of Computer Science, Jamia Millia Islamia New Delhi, India shakilkashish@yahoo.co.in, shadamanis@gmail.com and malam2@jmi.ac.in, https://pdfs.semanticscholar.org/1d1c/8ea500ca91038d6d43e337ef025bafb0bbda.pdf

9. EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES Aditya Methaila1 , Prince Kansal2 , Himanshu Arya3 , Pankaj Kumar4 1Netaji Subhas Institute of Technology,India and 2 Student, B.Tech (CSE), Maharaja Surajmal Institute of Technology New Delhi, India, https://airccj.org/CSCP/vol4/csit42607.pdf.

10. Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques, International Journal of Pure and Applied Mathematics, Volume 118 No. 8 2018, 165-174 ISSN: 1311-8080 , ISSN: 1314-3395 (on-line version) url: http://www.ijpam.eu.

11. Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014, https://pdfs.semanticscholar.org/2456/a979fbe8eea47b90d625c1a064162be5382e.pdf https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/IntroductionToWeka.pdf

12. Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms Sanjay Kumar Sen Asst. Professor, Computer Science & Engg. Orissa Engineering College, Bhubaneswar, Odisha – India, International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 6 Issue 6 June 2017, Page No. 21623-21631 Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.14.