

**American International University-Bangladesh (AIUB)**

##### 

##### **DATA WAREHOUSING AND DATA MINING [CS]**

**PROJECTS:**

1: Supervised Learning

2: unsupervised Learning

**NAME:** TAREQ, MD IKRAM

**ID :** 16-32231-2

**SEC :** B

**Project-Supervised**

**Introduction:** Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value.

**Datasets: Echocardiogram Data** Set had been used for performing supervised learning.

**Dataset Information:** The data for classifying if patients will survive for at least one year after a heart attack. The dataset was donated by Dr. Steven Salzberg which was collected by Dr. Evlin Kinney. The attribute characteristics of the data is { Categorical,Integer,Real }.

**Others Information are:**

Number of Instances: 132

Number of Attributes: 13

Missing Values: Yes

**Attribute Information:**

1. Survival -- the number of month’s patient survived.
2. Still-alive -- a binary variable. 0=dead at end of survival period, 1 means still alive.
3. Age-at-heart-attack -- age in years when heart attack occurred.
4. Pericardial-effusion -- binary. Pericardial effusion is fluid around the heart. 0=no fluid, 1=fluid.
5. Fractional-shortening -- a measure of contractility around the heart lower numbers are increasingly abnormal.
6. Epps -- E-point septal separation, another measure of contractility. Larger numbers are increasingly abnormal.
7. Ivdd -- left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
8. Wall-motion-score -- a measure of how the segments of the left ventricle are moving.
9. Wall-motion-index -- equals wall-motion-score divided by number of segments seen.
10. Mult -- a derivate var which can be ignored.
11. Name -- the name of the patient Which is specified by “name” .
12. Group – meaningless.
13. Alive-at-1 -- Boolean-valued. Derived from the first two attributes. 0 means patient was either dead after 1 year or had been followed for less than 1 year. 1 means patient was alive at 1 year.

**Methodology:**

This dataset downloaded from [http://archive.ics.uci.edu/ml/datasets/Echocardiogram] the dataset was converted to .csv to insert into WEKA software.

The classifiers that have been used are:

1. Naïve Bayes

2. LMT

3. AttributeSelectedClassifier

4. LWL

5. VotedPerceptron

**RESULTS:**

**1.NaiveBayes:**

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: Echocardiogram\_Data

Instances: 132

Attributes: 13

survival

still\_alive

age\_at\_heart\_attack

pericardial\_effusion

fractional\_shortening

epss

lvdd

wall\_motion\_score

wall\_motion\_index

mult

name

group

alive\_at\_1

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute 0 1

(0.67) (0.33)

========================================

survival

mean 30.9861 2.1145

std. dev. 13.714 2.5215

weight sum 50 23

precision 1.3895 1.3895

still\_alive

0 46.0 1.0

1 6.0 25.0

[total] 52.0 26.0

age\_at\_heart\_attack

mean 62.8571 67.5776

std. dev. 7.9796 8.2919

weight sum 50 23

precision 1.4286 1.4286

pericardial\_effusion

0 44.0 17.0

1 8.0 9.0

[total] 52.0 26.0

fractional\_shortening

mean 0.2418 0.1715

std. dev. 0.114 0.0807

weight sum 48 22

precision 0.0158 0.0158

epss

mean 11.0617 15.873

std. dev. 5.8689 8.6854

weight sum 45 21

precision 0.7407 0.7407

lvdd

mean 4.6432 5.1112

std. dev. 0.714 0.773

weight sum 48 20

precision 0.0552 0.0552

wall\_motion\_score

mean 13.5806 18.795

std. dev. 3.7547 7.3011

weight sum 50 23

precision 0.9714 0.9714

wall\_motion\_index

mean 1.2779 1.7926

std. dev. 0.3361 0.473

weight sum 50 23

precision 0.0513 0.0513

mult

mean 0.8004 0.745

std. dev. 0.1773 0.1658

weight sum 50 23

precision 0.0308 0.0308

name

name 51.0 25.0

[total] 51.0 25.0

group

1 16.0 10.0

2 36.0 16.0

[total] 52.0 26.0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 71 95.9459 %

Incorrectly Classified Instances 3 4.0541 %

Kappa statistic 0.9065

Mean absolute error 0.0538

Root mean squared error 0.1904

Relative absolute error 12.2058 %

Root relative squared error 40.6201 %

Total Number of Instances 74

Ignored Class Unknown Instances 58

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.980 0.083 0.961 0.980 0.970 0.907 0.689 0.489 0

0.917 0.020 0.957 0.917 0.936 0.907 0.963 0.845 1

Weighted Avg. 0.959 0.063 0.959 0.959 0.959 0.907 0.778 0.605

=== Confusion Matrix ===

a b <-- classified as

49 1 | a = 0

2 22 | b = 1

**2.LMT:**

=== Run information ===

Scheme: weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0

Relation: Echocardiogram\_Data

Instances: 132

Attributes: 13

survival

still\_alive

age\_at\_heart\_attack

pericardial\_effusion

fractional\_shortening

epss

lvdd

wall\_motion\_score

wall\_motion\_index

mult

name

group

alive\_at\_1

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Logistic model tree

------------------

: LM\_1:6/6 (74)

Number of Leaves : 1

Size of the Tree : 1

LM\_1:

Class 0 :

-2.36 +

[survival] \* 0.11 +

[still\_alive=1] \* -2.35 +

[lvdd] \* 0.54

Class 1 :

2.36 +

[survival] \* -0.11 +

[still\_alive=1] \* 2.35 +

[lvdd] \* -0.54

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 70 94.5946 %

Incorrectly Classified Instances 4 5.4054 %

Kappa statistic 0.8767

Mean absolute error 0.0576

Root mean squared error 0.1579

Relative absolute error 13.0724 %

Root relative squared error 33.6989 %

Total Number of Instances 74

Ignored Class Unknown Instances 58

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.960 0.083 0.960 0.960 0.960 0.877 0.726 0.650 0

0.917 0.040 0.917 0.917 0.917 0.877 0.942 0.654 1

Weighted Avg. 0.946 0.069 0.946 0.946 0.946 0.877 0.796 0.651

=== Confusion Matrix ===

a b <-- classified as

48 2 | a = 0

2 22 | b = 1

**3.AttributeSelectedClassifier:**

=== Run information ===

Scheme: weka.classifiers.meta.AttributeSelectedClassifier -E "weka.attributeSelection.CfsSubsetEval -P 1 -E 1" -S "weka.attributeSelection.BestFirst -D 1 -N 5" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Relation: Echocardiogram\_Data

Instances: 132

Attributes: 13

survival

still\_alive

age\_at\_heart\_attack

pericardial\_effusion

fractional\_shortening

epss

lvdd

wall\_motion\_score

wall\_motion\_index

mult

name

group

alive\_at\_1

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

AttributeSelectedClassifier:

=== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 67

Merit of best subset found: 0.859

Attribute Subset Evaluator (supervised, Class (nominal): 13 alive\_at\_1):

CFS Subset Evaluator

Including locally predictive attributes

Selected attributes: 1,2,4,9 : 4

survival

still\_alive

pericardial\_effusion

wall\_motion\_index

Header of reduced data:

@relation 'Echocardiogram\_Data-weka.filters.unsupervised.attribute.Remove-V-R1-2,4,9,13'

@attribute survival numeric

@attribute still\_alive {0,1}

@attribute pericardial\_effusion {0,1}

@attribute wall\_motion\_index numeric

@attribute alive\_at\_1 {0,1}

@data

Classifier Model

J48 pruned tree

------------------

survival <= 10: 1 (24.33/1.0)

survival > 10: 0 (49.67/0.67)

Number of Leaves : 2

Size of the tree : 3

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 72 97.2973 %

Incorrectly Classified Instances 2 2.7027 %

Kappa statistic 0.937

Mean absolute error 0.0434

Root mean squared error 0.1417

Relative absolute error 9.8567 %

Root relative squared error 30.2428 %

Total Number of Instances 74

Ignored Class Unknown Instances 58

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 0.083 0.962 1.000 0.980 0.939 0.701 0.505 0

0.917 0.000 1.000 0.917 0.957 0.939 0.926 0.609 1

Weighted Avg. 0.973 0.056 0.974 0.973 0.973 0.939 0.774 0.539

=== Confusion Matrix ===

a b <-- classified as

50 0 | a = 0

2 22 | b = 1

**4.LWL:**

=== Run information ===

Scheme: weka.classifiers.lazy.LWL -U 0 -K -1 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\"" -W weka.classifiers.trees.DecisionStump

Relation: Echocardiogram\_Data

Instances: 132

Attributes: 13

survival

still\_alive

age\_at\_heart\_attack

pericardial\_effusion

fractional\_shortening

epss

lvdd

wall\_motion\_score

wall\_motion\_index

mult

name

group

alive\_at\_1

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Locally weighted learning

===========================

Using classifier: weka.classifiers.trees.DecisionStump

Using linear weighting kernels

Using all neighbours

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 71 95.9459 %

Incorrectly Classified Instances 3 4.0541 %

Kappa statistic 0.9065

Mean absolute error 0.0472

Root mean squared error 0.1797

Relative absolute error 10.7146 %

Root relative squared error 38.3441 %

Total Number of Instances 74

Ignored Class Unknown Instances 58

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.980 0.083 0.961 0.980 0.970 0.907 0.745 0.561 0

0.917 0.020 0.957 0.917 0.936 0.907 0.934 0.689 1

Weighted Avg. 0.959 0.063 0.959 0.959 0.959 0.907 0.806 0.603

=== Confusion Matrix ===

a b <-- classified as

49 1 | a = 0

2 22 | b = 1

**5.VotedPerceptron:**

=== Run information ===

Scheme: weka.classifiers.functions.VotedPerceptron -I 1 -E 1.0 -S 1 -M 10000

Relation: Echocardiogram\_Data

Instances: 132

Attributes: 13

survival

still\_alive

age\_at\_heart\_attack

pericardial\_effusion

fractional\_shortening

epss

lvdd

wall\_motion\_score

wall\_motion\_index

mult

name

group

alive\_at\_1

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

VotedPerceptron: Number of perceptrons=14

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 69 93.2432 %

Incorrectly Classified Instances 5 6.7568 %

Kappa statistic 0.8407

Mean absolute error 0.0702

Root mean squared error 0.2427

Relative absolute error 15.9504 %

Root relative squared error 51.7877 %

Total Number of Instances 74

Ignored Class Unknown Instances 58

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.980 0.167 0.925 0.980 0.951 0.845 0.708 0.515 0

0.833 0.020 0.952 0.833 0.889 0.845 0.925 0.630 1

Weighted Avg. 0.932 0.119 0.934 0.932 0.931 0.845 0.778 0.552

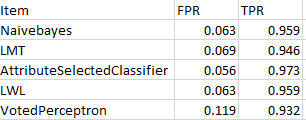
=== Confusion Matrix ===

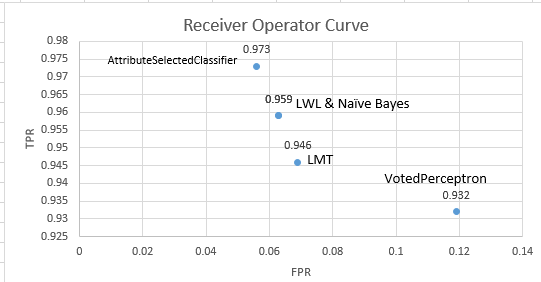
a b <-- classified as

49 1 | a = 0

4 20 | b = 1

**RESULT:**





**Analysis:**

We can see from the analysis that AttributeSelectedClassifier has the highest Correctly Classified Instances which is 97.3%. Here from the ROC graph we can see AttributeSelectedClassifier is the nearest to the best point of the ROC graph. So AttributeSelectedClassifier can be chosen as the best classifier.