**Dataset Name:** Echocardiogram

**Problem:**

All the patients suffered heart attacks at some point in the past. Some are still alive and some are not. The survival and still-alive variables, when taken together, indicate whether a patient survived for at least one year following the heart attack.

The problem addressed by past researchers was to predict from the other variables whether or not the patient will survive at least one year.

**Objective:**

* Analyzing and comparing performance of the following classifier using ROC Graph:

1. IBK
2. Decision Stamp
3. J48
4. Naive Bayes
5. KStar

* Finding the most suitable classifier.

**Dataset Prepare:**

In this dataset, there are a lot of missing values. My task was to determine whether the patient will survive one year or not. For that I have chosen alive-at-1 attribute as my decision class and the missing values handled for both Still-alive and survival attribute value was missing than I decided to discard that instance and for other attribute missing value was filled by applying DMI into the data set .I ignore alive-at-1 when I was applying DMI into the dataset.

For alive-at-1 attribute , we can see that for all still-alive’s 0 value, alive-at-1 value is 0, so that I filled up all missing value of alive-at-1 with 0 for having still-alive’s 0. For alive-at-1’s value 1, I have seen that survival < 12 months and age-at-heart-attack’s value is >=60, they survived. So, I filled up those cases missing values with 1 and for the rest of the cases I have filled up with 0.

**1.IBK Classifier:**

Test mode: 131-fold cross-validation

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

IB1 instance-based classifier

using 1 nearest neighbour(s) for classification

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 123 93.8931 %

Incorrectly Classified Instances 8 6.1069 %

Kappa statistic 0.8495

Mean absolute error 0.0677

Root mean squared error 0.2454

Relative absolute error 17.0924 %

Root relative squared error 55.0315 %

Total Number of Instances 131

=== Detailed Accuracy By Class ===

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

0.943 0.063 0.846 0.943 0.892 0.852 0.940 0.813 1

0.938 0.057 0.978 0.938 0.957 0.852 0.940 0.963 0

Weighted Avg. 0.939 0.059 0.943 0.939 0.940 0.852 0.940 0.923

=== Confusion Matrix ===

a b <-- classified as

33 2 | a = 1 (survive)

6 90 | b = 0 (not survive)

**2.Decision Stump Classifier:**

Test mode: 131-fold cross-validation

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

Decision Stump

Classifications A

survival <= 8.25 : 1

survival > 8.25 : 0

survival is missing : 0

Class distributions

survival <= 8.25

1 0

1.0 0.0

survival > 8.25

1 0

0.010309278350515464 0.9896907216494846

survival is missing

1 0

0.26717557251908397 0.732824427480916

=== Summary ===

Correctly Classified Instances 130 99.2366 %

Incorrectly Classified Instances 1 0.7634 %

Kappa statistic 0.9803

Mean absolute error 0.0153

Root mean squared error 0.0878

Relative absolute error 3.8535 %

Root relative squared error 19.6981 %

Total Number of Instances 131

=== Detailed Accuracy By Class ===

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

0.971 0.000 1.000 0.971 0.986 0.981 0.971 0.979 1

1.000 0.029 0.990 1.000 0.995 0.981 0.971 0.990 0

Weighted Avg. 0.992 0.021 0.992 0.992 0.992 0.981 0.971 0.987

=== Confusion Matrix ===

a b <-- classified as

34 1 | a = 1 (survive)

0 96 | b = 0 (not survive)

**3.J48 classifier:**

Test mode: 131-fold cross-validation

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

J48 pruned tree

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

survival <= 7.5: 1 (34.0)

survival > 7.5: 0 (97.0/1.0)

Number of Leaves : 2

Size of the tree : 3

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 129 98.4733 %

Incorrectly Classified Instances 2 1.5267 %

Kappa statistic 0.9603

Mean absolute error 0.0228

Root mean squared error 0.1232

Relative absolute error 5.7603 %

Root relative squared error 27.6434 %

Total Number of Instances 131

=== Detailed Accuracy By Class ===

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

0.943 0.000 1.000 0.943 0.971 0.961 0.943 0.958 1

1.000 0.057 0.980 1.000 0.990 0.961 0.943 0.980 0

Weighted Avg. 0.985 0.042 0.985 0.985 0.985 0.961 0.943 0.974

=== Confusion Matrix ===

a b <-- classified as

33 2 | a = 1 (survive)

0 96 | b = 0 (not survive)

**4.Naïve Bayes Classifier:**

Test mode: 131-fold cross-validation

=== Summary ===

Correctly Classified Instances 128 97.7099 %

Incorrectly Classified Instances 3 2.2901 %

Kappa statistic 0.9399

Mean absolute error 0.0239

Root mean squared error 0.1192

Relative absolute error 6.0226 %

Root relative squared error 26.7414 %

Total Number of Instances 131

=== Detailed Accuracy By Class ===

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

0.914 0.000 1.000 0.914 0.955 0.942 1.000 0.999 1

1.000 0.086 0.970 1.000 0.985 0.942 1.000 1.000 0

Weighted Avg. 0.977 0.063 0.978 0.977 0.977 0.942 1.000 1.000

=== Confusion Matrix ===

a b <-- classified as

32 3 | a = 1 (survive)

0 96 | b = 0 (not survive)

**5.KStar Classifier:**

Test mode: 131-fold cross-validation

=== Summary ===

Correctly Classified Instances 127 96.9466 %

Incorrectly Classified Instances 4 3.0534 %

Kappa statistic 0.922

Mean absolute error 0.0502

Root mean squared error 0.1691

Relative absolute error 12.6648 %

Root relative squared error 37.9311 %

Total Number of Instances 131

=== Detailed Accuracy By Class ===

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

0.943 0.021 0.943 0.943 0.943 0.922 0.990 0.979 1

0.979 0.057 0.979 0.979 0.979 0.922 0.990 0.996 0

Weighted Avg. 0.969 0.047 0.969 0.969 0.969 0.922 0.990 0.992

=== Confusion Matrix ===

a b <-- classified as

33 2 | a = 1 (survive)

2 94 | b = 0 (not survive)

**ROC Graph:**

|  |  |  |  |
| --- | --- | --- | --- |
| Classifiers | True Positive Rate (TPR) | False Positive Rate (FPR) | Linear Distance |
| IBK Classifier | 0.943 | 0.0625 | 0.085 |
| Decision Stump Classifier | 0.971 | 0 | 0.026 |
| J48 Classifier | 0.942 | 0 | 0.058 |
| Naive Bayes Classifier | 0.914 | 0 | 0.086 |
| KStar Classifier | 0.943 | 0.021 | 0.061 |

**Summery(Classifier performance):**

IBK classifier : IBK classifier FP is 6 .This classifier classify 6 as survive wrongfully among 96 not survive patient .

Decision Stump : Decision classifier FP is 0 .This classifier can full fill to our main concern to classify not survive patient accurately.

J48 : J48 classifier FP is also 0 but this classifier accuracy is little bit less than decision stump classifier .

Naïve Bayes classifier : Naïve Bayes classifier accuracy is less than decision stump and J48 classifier. Naïve Bayes classifier stayed far from best possible classifier in linear distance but it can classify not survive patient accurately for this reason this classifier is better than KStar and IBK classifier .

KStar classifier : KStar FP value is 2 .Which means this classifier wrongfully classify 2 as survive among 96 not survive patient .

So that, Decision Stump>J48> Naïve Bayes >KStar >IBK

**Comment (Best classifier):**

We can see in the ROC graph that Decision Stump is giving the best result for my interest. The FP rate is lower as it gave only 0 which means this classifier can predict not survive accurately .Decision stump classifier accuracy is higher among the given five classifier by this classifier can classify 130 instances among 131 instances. Decision Stump classifier also more closest with the best possible classifier point in ROC graph .

Sothat, for the echocardiogram dataset problem I want to choose decision stump classifier.