DATA MINING HOMEWORK with WEKA

by

İrem Yuvalı

2017555071

January, 2022

TABLE OF CONTENTS

PREFACE

The document contains the data analyses of two datasets according to the four classifier types named as Naive Bayes, J48, OneR and IBk. The mission of the project is to create and compare the analyses using by WEKA tool for CEN-481 Introduction to Data Mining Course given by Havva Esin Ünal.

1. DATASETS

1.1 THE FIRST DATASET: BREAST CANCER

Overview:

This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature.

This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal. **[1]**

Sources:

Matjaz Zwitter & Milan Soklic (physicians) Institute of Oncology University Medical Center Ljubljana, Yugoslavia

Donors: Ming Tan and Jeff Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)

Date: 11 July 1988

Past Usage:

- Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986). The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann.
    - accuracy range: 66%-72%

- Clark,P. & Niblett,T. (1987). Induction in Noisy Domains. In Progress in Machine Learning (from the Proceedings of the 2nd European Working Session on Learning), 11-30, Bled, Yugoslavia: Sigma Press.
    - 8 test results given: 65%-72% accuracy range

- Tan, M., & Eshelman, L. (1988). Using weighted networks to represent classification knowledge in noisy domains. Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI.
    - 4 systems tested: accuracy range was 68%-73.5%

- Cestnik,G., Konenenko,I, & Bratko,I. (1987). Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press.
    - Assistant-86: 78% accuracy

Attributes:

- Number of Instances: 286
- Number of Attributes: 9 + the class attribute

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

3. menopause: lt40, ge40, premeno.

4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.

5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.

6. node-caps: yes, no.

7. deg-malig: 1, 2, 3.

8. breast: left, right.

9. breast-quad: left-up, left-low, right-up, right-low, central.

10. irradiat: yes, no.

- Missing Attribute Values: (denoted by "?")
- Attribute #: Number of instances with missing values:

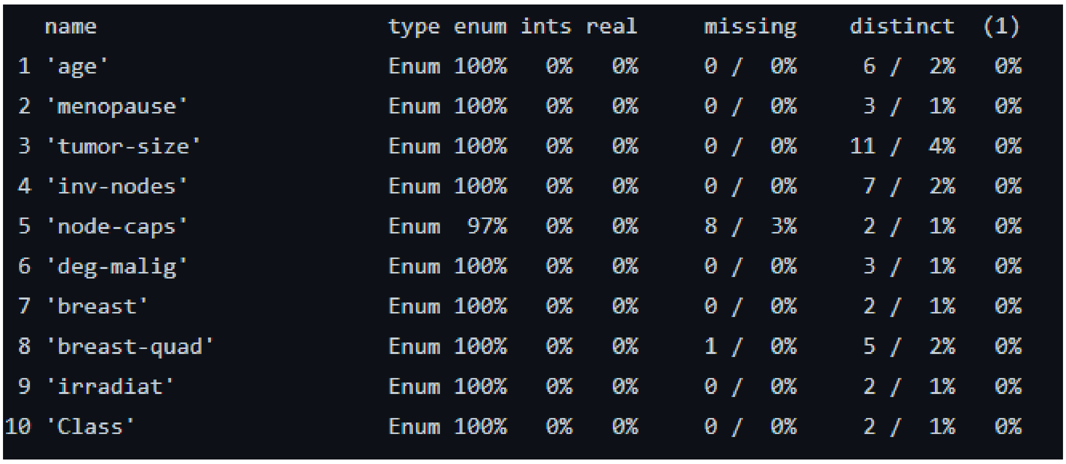
6.	8
9.	1.

- Class Distribution:

1. no-recurrence-events: 201 instances

2. recurrence-events: 85 instances

- Num Instances: 286
- Num Attributes: 10
- Num Continuous: 0 (Int 0 / Real 0)
- Num Discrete: 10
- Missing values: 9 / 0.3%



```
  name                     type enum ints real     missing    distinct  (1)
1 'age'                    Enum 100%   0%   0%     0 /  0%     6 /  2%   0%
2 'menopause'              Enum 100%   0%   0%     0 /  0%     3 /  1%   0%
3 'tumor-size'             Enum 100%   0%   0%     0 /  0%    11 /  4%   0%
4 'inv-nodes'              Enum 100%   0%   0%     0 /  0%     7 /  2%   0%
5 'node-caps'              Enum  97%   0%   0%     8 /  3%     2 /  1%   0%
6 'deg-malig'              Enum 100%   0%   0%     0 /  0%     3 /  1%   0%
7 'breast'                 Enum 100%   0%   0%     0 /  0%     2 /  1%   0%
8 'breast-quad'            Enum 100%   0%   0%     1 /  0%     5 /  2%   0%
9 'irradiat'               Enum 100%   0%   0%     0 /  0%     2 /  1%   0%
10 'Class'                 Enum 100%   0%   0%     0 /  0%     2 /  1%   0%
```

1.2 THE SECOND DATASET: HEART DISEASE

Overview:

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease. **[2]**

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Sources:

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations
- Total: 1190 observations
- Duplicated: 272 observations

Final dataset: 918 observations

Creators:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
- Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Attributes:

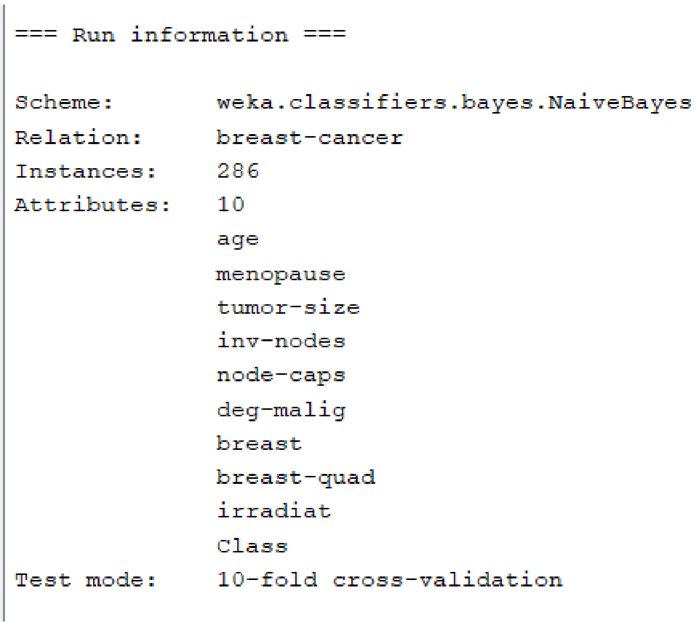
1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

2. ANALYSES

2.1 THE FIRST DATASET: BREAST CANCER

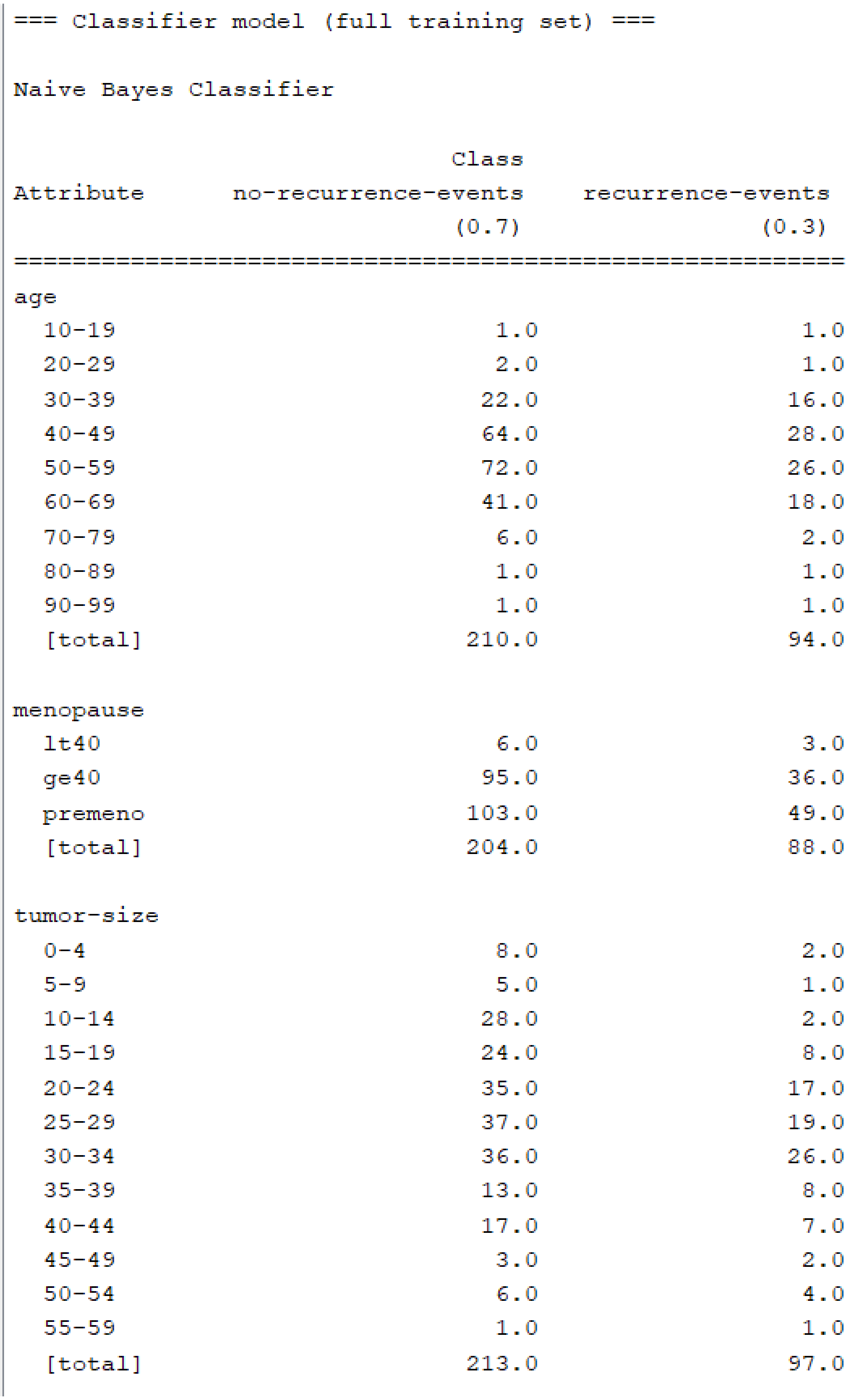
2.1.1 Naive Bayes (Naive Bayes Classifier)

Run Information:



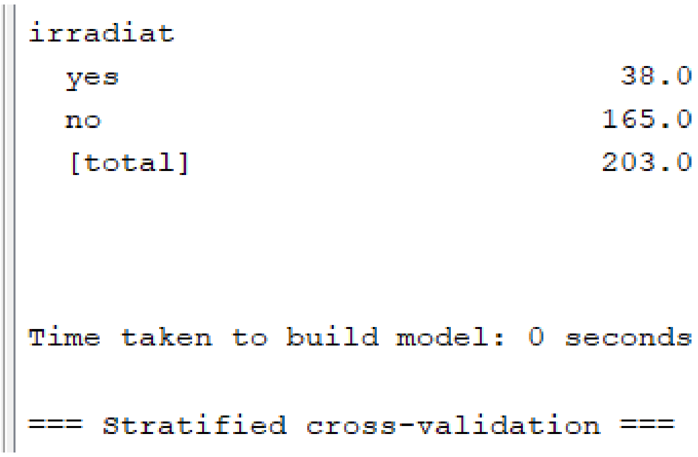
```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     breast-cancer
Instances:    286
Attributes:   10
              age
              menopause
              tumor-size
              inv-nodes
              node-caps
              deg-malig
              breast
              breast-quad
              irradiat
              Class
Test mode:    10-fold cross-validation
```

Training set:



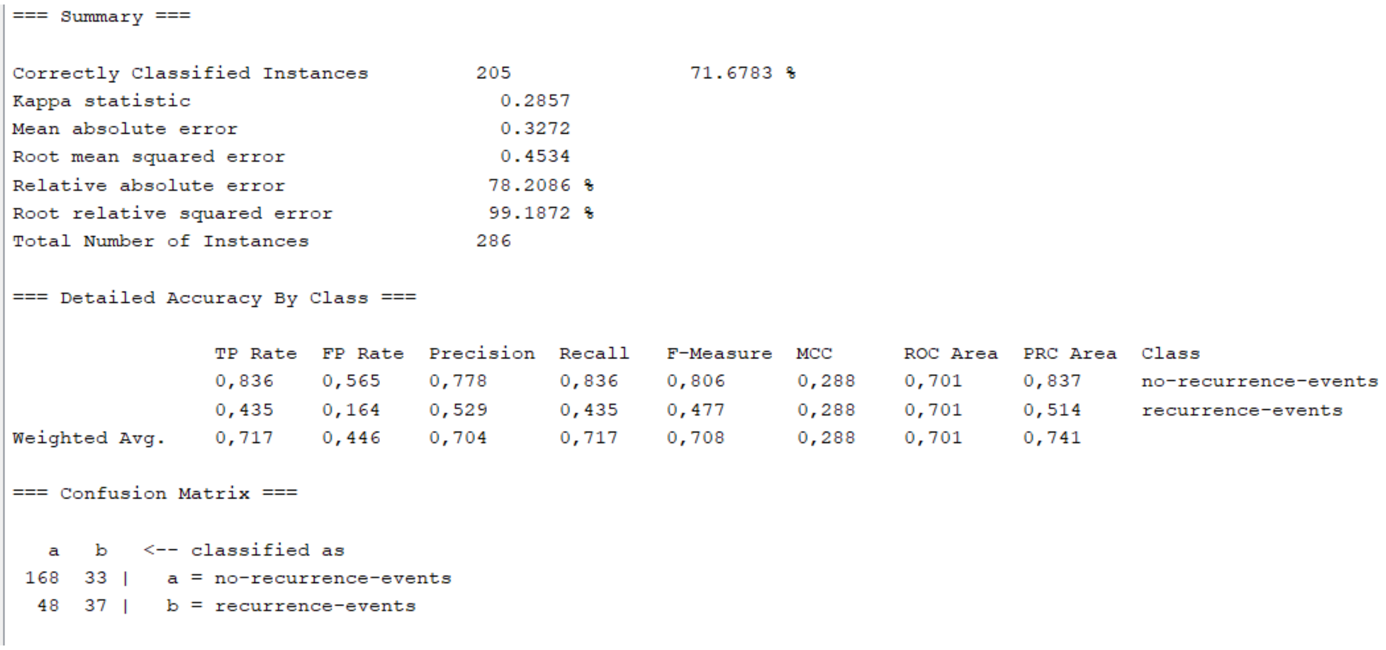
```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                                    Class
Attribute          no-recurrence-events    recurrence-events
                                    (0.7)                (0.3)
=========================================================
age
  10-19                               1.0                  1.0
  20-29                               2.0                  1.0
  30-39                              22.0                 16.0
  40-49                              64.0                 28.0
  50-59                              72.0                 26.0
  60-69                              41.0                 18.0
  70-79                               6.0                  2.0
  80-89                               1.0                  1.0
  90-99                               1.0                  1.0
  [total]                           210.0                 94.0

menopause
  lt40                                6.0                  3.0
  ge40                               95.0                 36.0
  premeno                           103.0                 49.0
  [total]                           204.0                 88.0

tumor-size
  0-4                                 8.0                  2.0
  5-9                                 5.0                  1.0
  10-14                              28.0                  2.0
  15-19                              24.0                  8.0
  20-24                              35.0                 17.0
  25-29                              37.0                 19.0
  30-34                              36.0                 26.0
  35-39                              13.0                  8.0
  40-44                              17.0                  7.0
  45-49                               3.0                  2.0
  50-54                               6.0                  4.0
  55-59                               1.0                  1.0
  [total]                           213.0                 97.0
```

```
inv-nodes
  0-2                          168.0                 47.0
  3-5                           20.0                 18.0
  6-8                            8.0                 11.0
  9-11                           5.0                  7.0
  12-14                          2.0                  3.0
  15-17                          4.0                  4.0
  18-20                          1.0                  1.0
  21-23                          1.0                  1.0
  24-26                          1.0                  2.0
  27-29                          1.0                  1.0
  30-32                          1.0                  1.0
  33-35                          1.0                  1.0
  36-39                          1.0                  1.0
  [total]                      214.0                 98.0

node-caps
  yes                           26.0                 32.0
  no                           172.0                 52.0
  [total]                      198.0                 84.0

deg-malig
  1                             60.0                 13.0
  2                            103.0                 29.0
  3                             41.0                 46.0
  [total]                      204.0                 88.0

breast
  left                         104.0                 50.0
  right                         99.0                 37.0
  [total]                      203.0                 87.0

breast-quad
  left_up                       72.0                 27.0
  left_low                      76.0                 36.0
  right_up                      21.0                 14.0
  right_low                     19.0                  7.0
  central                       18.0                  5.0
  [total]                      206.0                 89.0
```



```
irradiat
  yes                        38.0               32.0
  no                        165.0               55.0
  [total]                   203.0               87.0



Time taken to build model: 0 seconds

=== Stratified cross-validation ===
```

Summary:

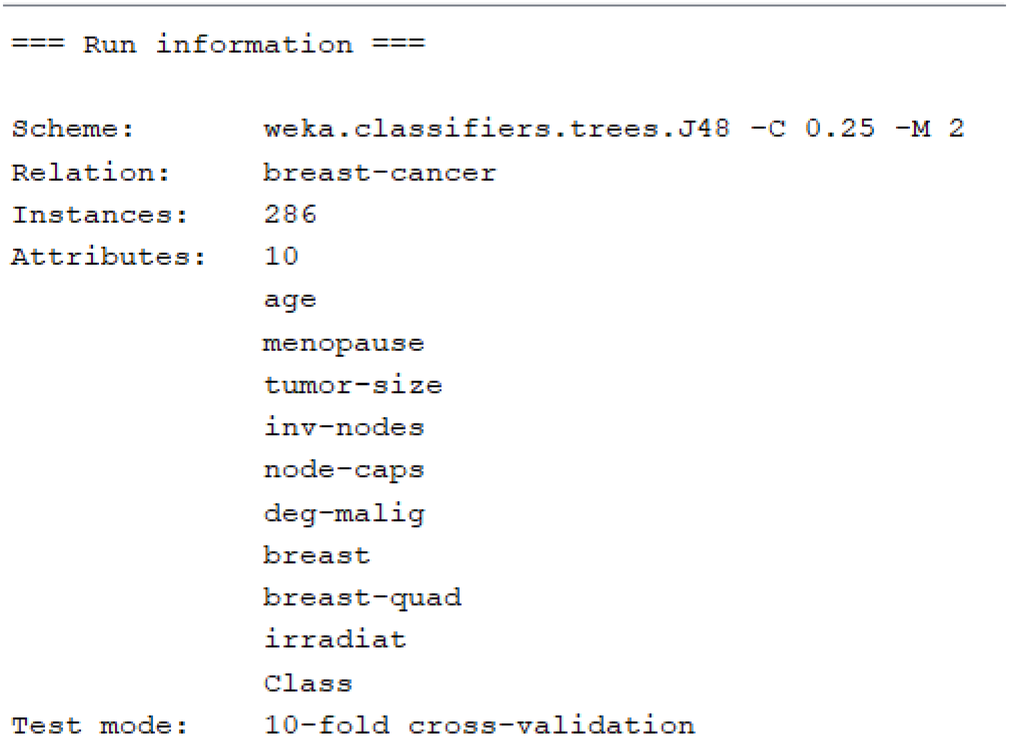


```
=== Summary ===

Correctly Classified Instances         205               71.6783 %
Kappa statistic                          0.2857
Mean absolute error                      0.3272
Root mean squared error                  0.4534
Relative absolute error                 78.2086 %
Root relative squared error             99.1872 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,836    0,565    0,778      0,836    0,806      0,288    0,701     0,837     no-recurrence-events
                 0,435    0,164    0,529      0,435    0,477      0,288    0,701     0,514     recurrence-events
Weighted Avg.    0,717    0,446    0,704      0,717    0,708      0,288    0,701     0,741

=== Confusion Matrix ===

   a   b   <-- classified as
 168  33 |   a = no-recurrence-events
  48  37 |   b = recurrence-events
```

According to the summary analysis, the model has achieved an accuracy of 71.6783% in the test set. From the total 286 instances, 205 of them are correctly classified.

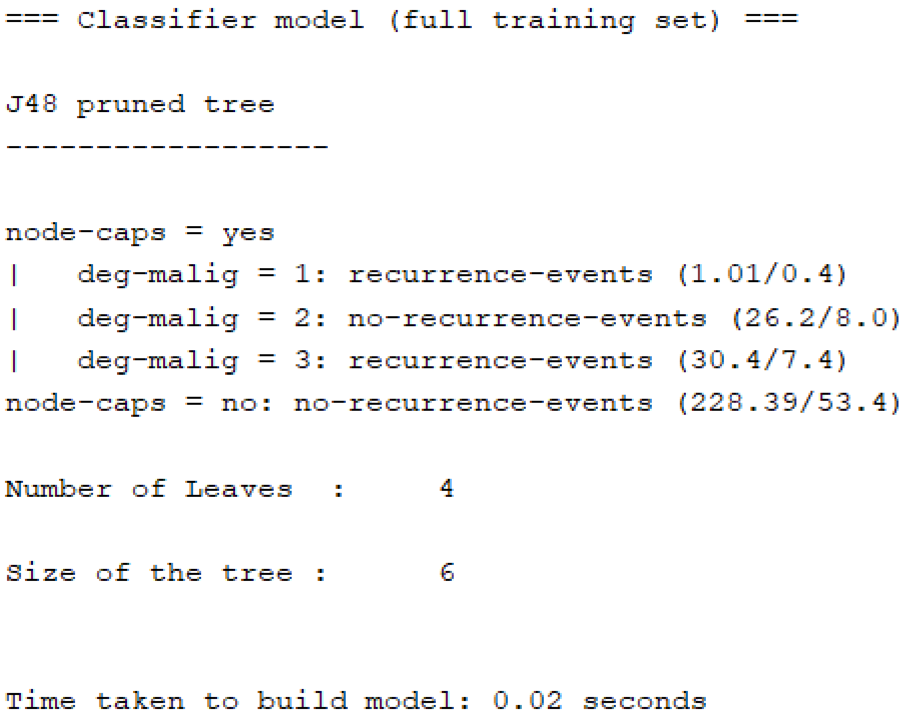
2.1.2 J48 (Decision Tree Classifier)

Run Information:



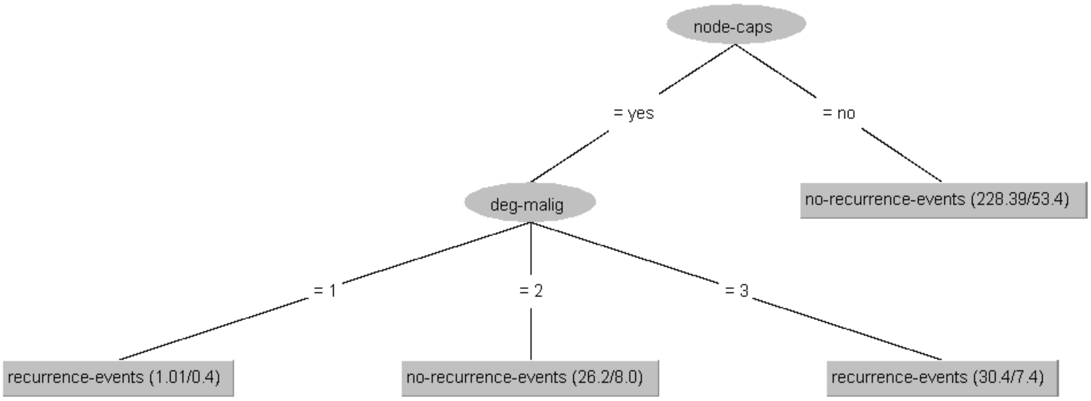
```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     breast-cancer
Instances:    286
Attributes:   10
              age
              menopause
              tumor-size
              inv-nodes
              node-caps
              deg-malig
              breast
              breast-quad
              irradiat
              Class
Test mode:    10-fold cross-validation
```

Training Set:

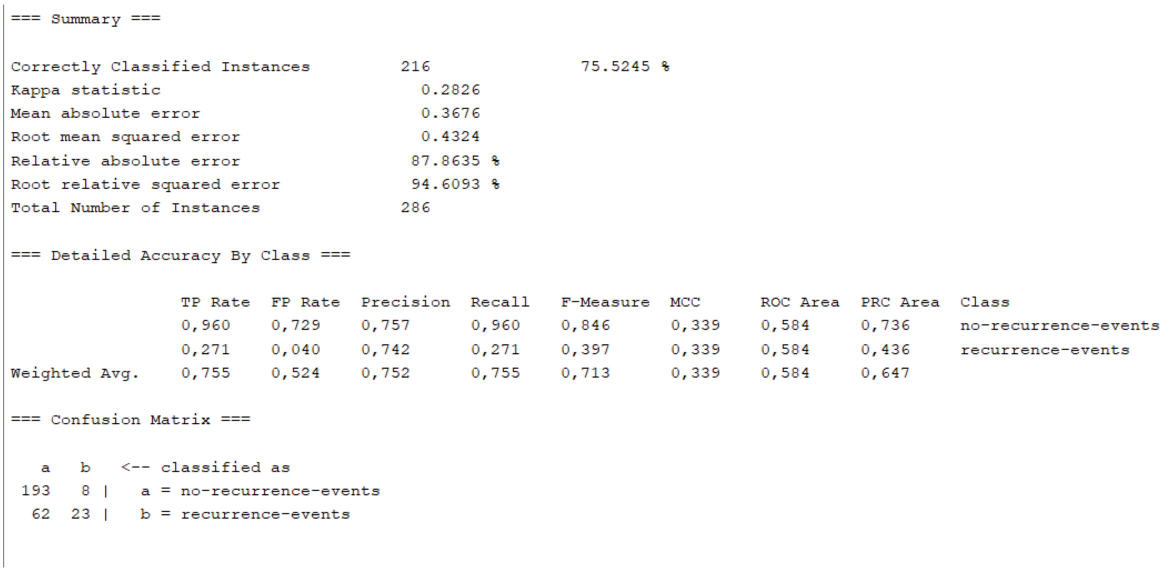


```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

node-caps = yes
|   deg-malig = 1: recurrence-events (1.01/0.4)
|   deg-malig = 2: no-recurrence-events (26.2/8.0)
|   deg-malig = 3: recurrence-events (30.4/7.4)
node-caps = no: no-recurrence-events (228.39/53.4)

Number of Leaves  :     4

Size of the tree :      6


Time taken to build model: 0.02 seconds
```

Tree Version:


node-caps
= yes
= no
no-recurrence-events (228.39/53.4)
deg-malig
= 1
= 2
= 3
recurrence-events (1.01/0.4)
no-recurrence-events (26.2/8.0)
recurrence-events (30.4/7.4)

Summary:

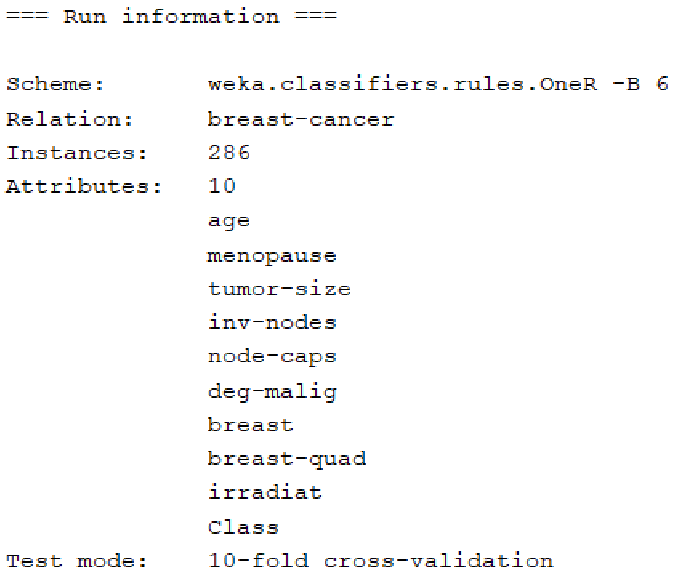


```
=== Summary ===

Correctly Classified Instances         216               75.5245 %
Kappa statistic                          0.2826
Mean absolute error                      0.3676
Root mean squared error                  0.4324
Relative absolute error                 87.8635 %
Root relative squared error             94.6093 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,960    0,729    0,757      0,960    0,846      0,339    0,584     0,736     no-recurrence-events
                 0,271    0,040    0,742      0,271    0,397      0,339    0,584     0,436     recurrence-events
Weighted Avg.    0,755    0,524    0,752      0,755    0,713      0,339    0,584     0,647

=== Confusion Matrix ===

   a   b   <-- classified as
 193   8 |   a = no-recurrence-events
  62  23 |   b = recurrence-events
```

According to the summary analysis, the model has achieved an accuracy of 75.5245% in the test set. From the total 286 instances, 216 of them are correctly classified.

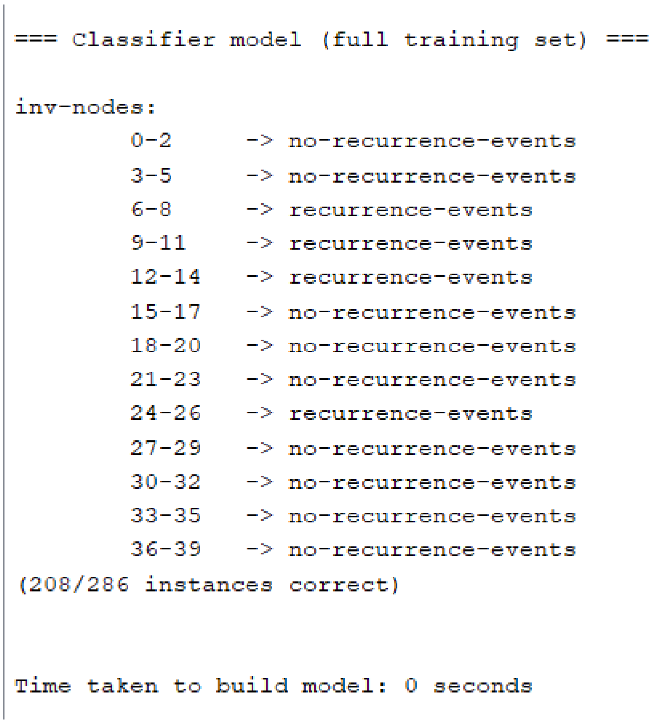
2.1.3 OneR (Rule-based Classifier)

Run Information:



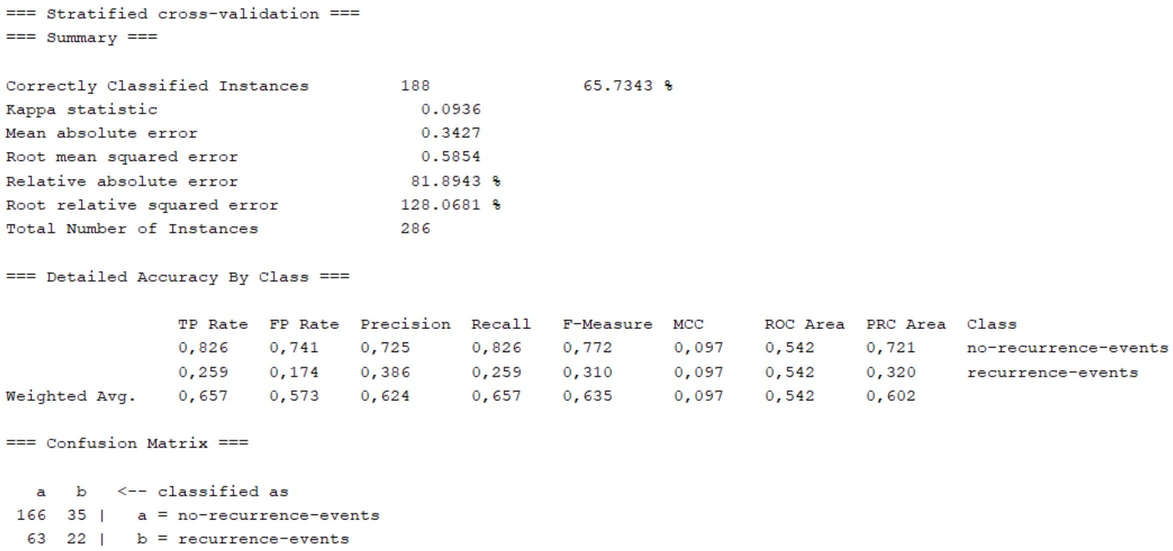
```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     breast-cancer
Instances:    286
Attributes:   10
              age
              menopause
              tumor-size
              inv-nodes
              node-caps
              deg-malig
              breast
              breast-quad
              irradiat
              Class
Test mode:    10-fold cross-validation
```

Training set:



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

inv-nodes:
	0-2	-> no-recurrence-events
	3-5	-> no-recurrence-events
	6-8	-> recurrence-events
	9-11	-> recurrence-events
	12-14	-> recurrence-events
	15-17	-> no-recurrence-events
	18-20	-> no-recurrence-events
	21-23	-> no-recurrence-events
	24-26	-> recurrence-events
	27-29	-> no-recurrence-events
	30-32	-> no-recurrence-events
	33-35	-> no-recurrence-events
	36-39	-> no-recurrence-events
(208/286 instances correct)


Time taken to build model: 0 seconds
```

Summary:

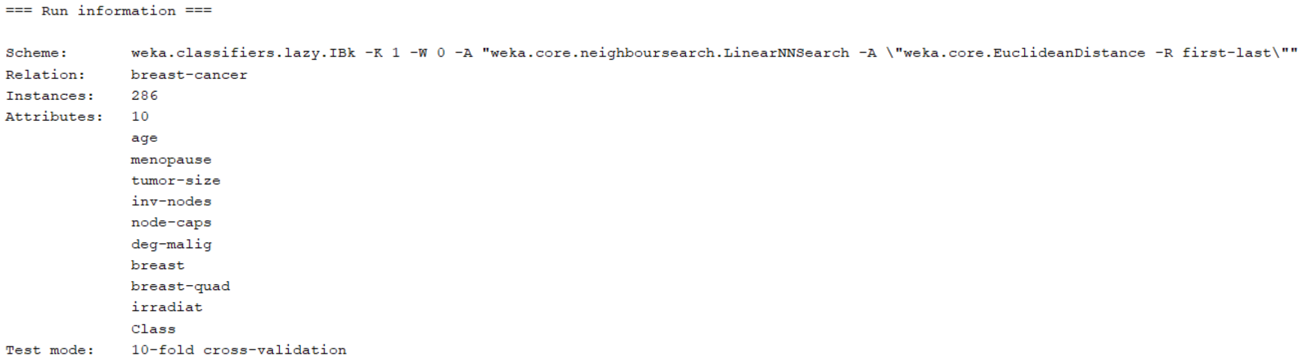


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         188               65.7343 %
Kappa statistic                          0.0936
Mean absolute error                      0.3427
Root mean squared error                  0.5854
Relative absolute error                 81.8943 %
Root relative squared error            128.0681 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,826    0,741    0,725      0,826    0,772      0,097    0,542     0,721     no-recurrence-events
                 0,259    0,174    0,386      0,259    0,310      0,097    0,542     0,320     recurrence-events
Weighted Avg.    0,657    0,573    0,624      0,657    0,635      0,097    0,542     0,602

=== Confusion Matrix ===

   a   b   <-- classified as
 166  35 |   a = no-recurrence-events
  63  22 |   b = recurrence-events
```

According to the summary analysis, the model has achieved an accuracy of 65.7343% in the test set. From the total 286 instances, 188 of them are correctly classified.

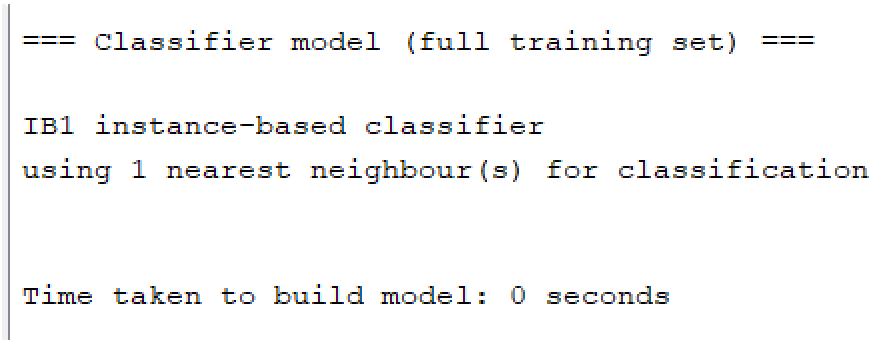
2.1.4 IBk (k-Nearest Neighbor Classifier)

Run Information:



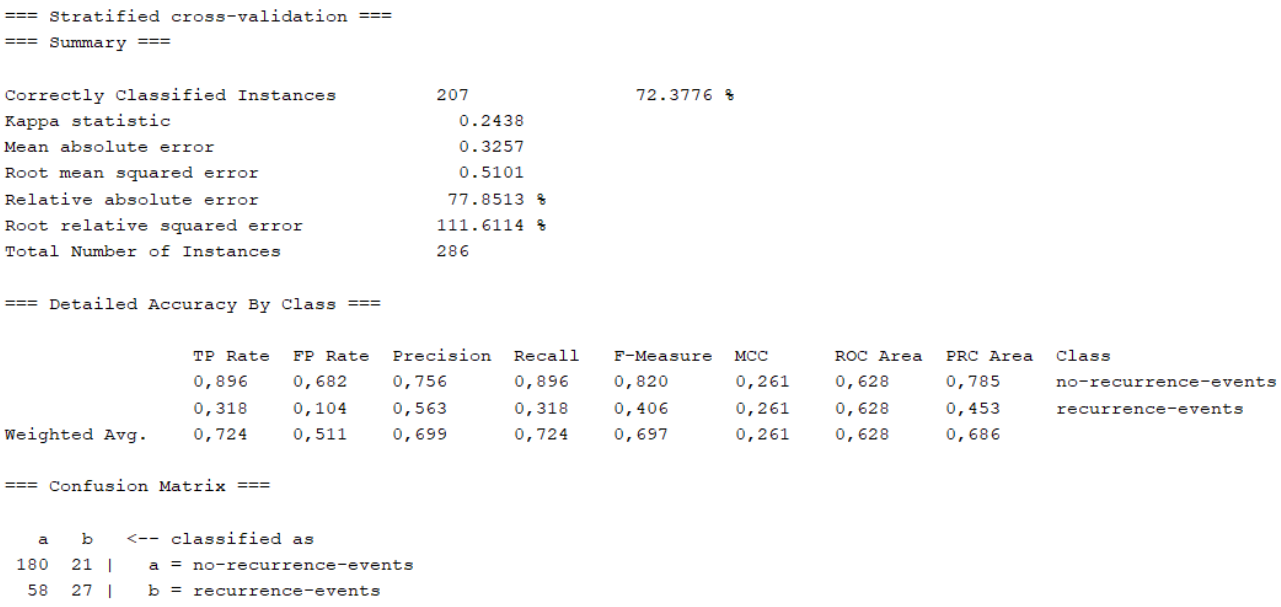
```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     breast-cancer
Instances:    286
Attributes:   10
              age
              menopause
              tumor-size
              inv-nodes
              node-caps
              deg-malig
              breast
              breast-quad
              irradiat
              Class
Test mode:    10-fold cross-validation
```

Training set:



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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds
```

Summary:



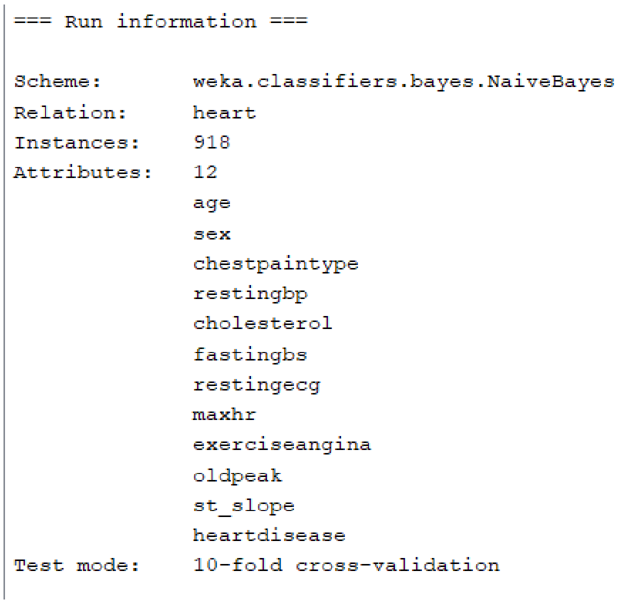
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         207               72.3776 %
Kappa statistic                          0.2438
Mean absolute error                      0.3257
Root mean squared error                  0.5101
Relative absolute error                 77.8513 %
Root relative squared error            111.6114 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,896    0,682    0,756      0,896    0,820      0,261    0,628     0,785     no-recurrence-events
                 0,318    0,104    0,563      0,318    0,406      0,261    0,628     0,453     recurrence-events
Weighted Avg.    0,724    0,511    0,699      0,724    0,697      0,261    0,628     0,686

=== Confusion Matrix ===

   a   b   <-- classified as
 180  21 |   a = no-recurrence-events
  58  27 |   b = recurrence-events
```

According to the summary analysis, the model has achieved an accuracy of 72.3776% in the test set. From the total 286 instances, 207 of them are correctly classified.

2.2 THE SECOND DATASET: HEART DISEASE

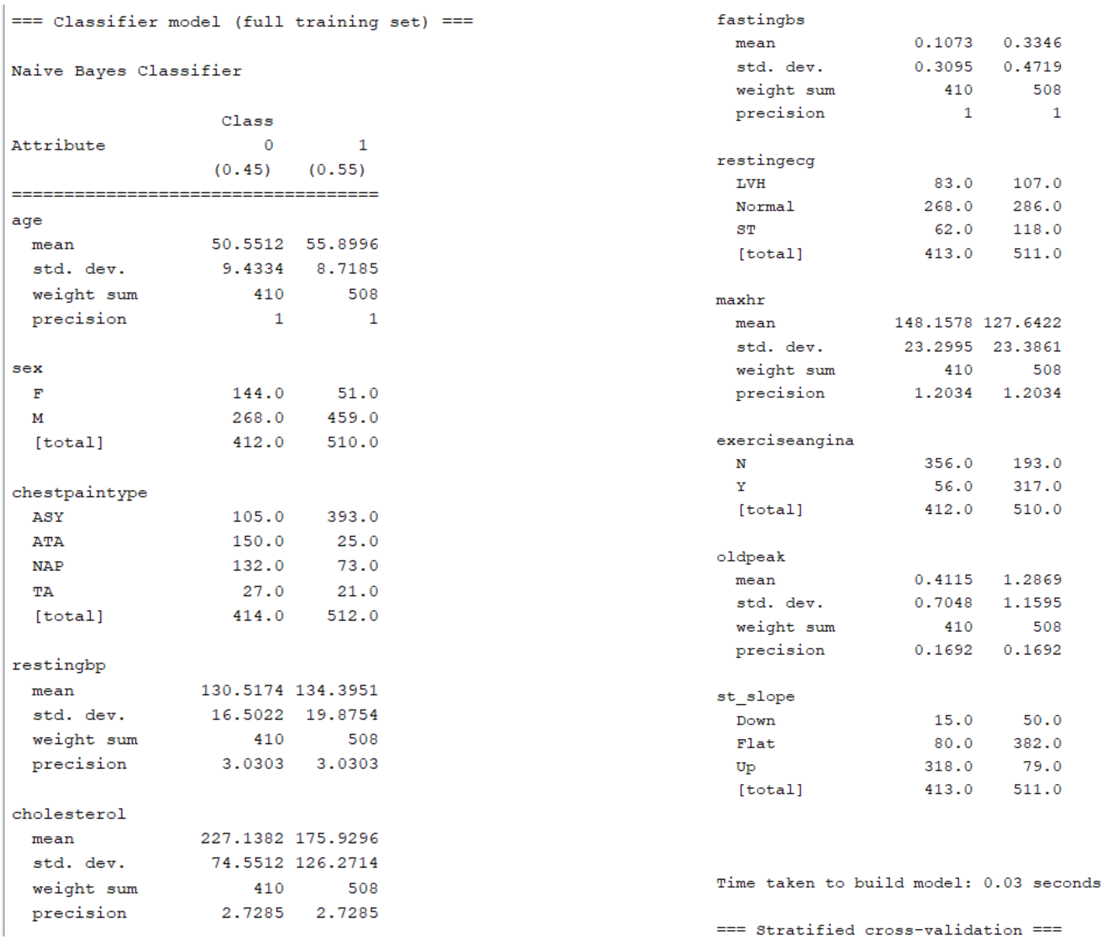
2.2.1 Naive Bayes (Naive Bayes Classifier)

Run Information:



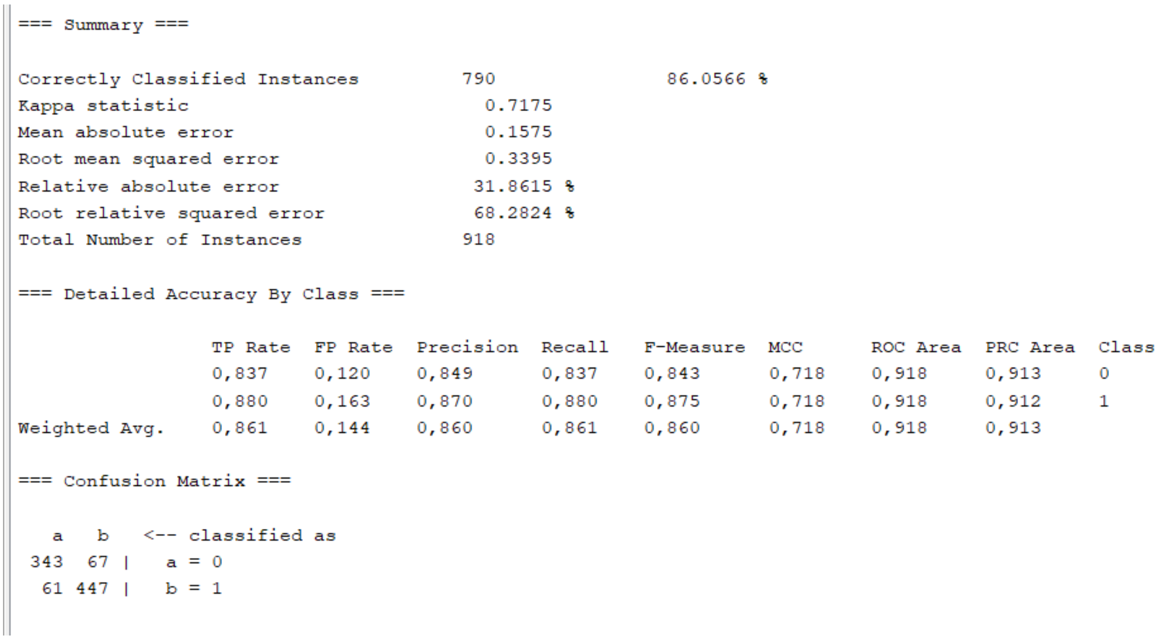
```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     heart
Instances:    918
Attributes:   12
              age
              sex
              chestpaintype
              restingbp
              cholesterol
              fastingbs
              restingecg
              maxhr
              exerciseangina
              oldpeak
              st_slope
              heartdisease
Test mode:    10-fold cross-validation
```

Training set:



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

Naive Bayes Classifier

                     Class
Attribute                0        1
                    (0.45)   (0.55)
===================================
age
  mean              50.5512  55.8996
  std. dev.          9.4334   8.7185
  weight sum            410      508
  precision               1        1

sex
  F                   144.0     51.0
  M                   268.0    459.0
  [total]             412.0    510.0

chestpaintype
  ASY                 105.0    393.0
  ATA                 150.0     25.0
  NAP                 132.0     73.0
  TA                   27.0     21.0
  [total]             414.0    512.0

restingbp
  mean             130.5174 134.3951
  std. dev.         16.5022  19.8754
  weight sum            410      508
  precision          3.0303   3.0303

cholesterol
  mean             227.1382 175.9296
  std. dev.         74.5512 126.2714
  weight sum            410      508
  precision          2.7285   2.7285

fastingbs
  mean               0.1073   0.3346
  std. dev.          0.3095   0.4719
  weight sum            410      508
  precision               1        1

restingecg
  LVH                  83.0    107.0
  Normal              268.0    286.0
  ST                   62.0    118.0
  [total]             413.0    511.0

maxhr
  mean             148.1578 127.6422
  std. dev.         23.2995  23.3861
  weight sum            410      508
  precision          1.2034   1.2034

exerciseangina
  N                   356.0    193.0
  Y                    56.0    317.0
  [total]             412.0    510.0

oldpeak
  mean               0.4115   1.2869
  std. dev.          0.7048   1.1595
  weight sum            410      508
  precision          0.1692   0.1692

st_slope
  Down                 15.0     50.0
  Flat                 80.0    382.0
  Up                  318.0     79.0
  [total]             413.0    511.0



Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
```

Summary:

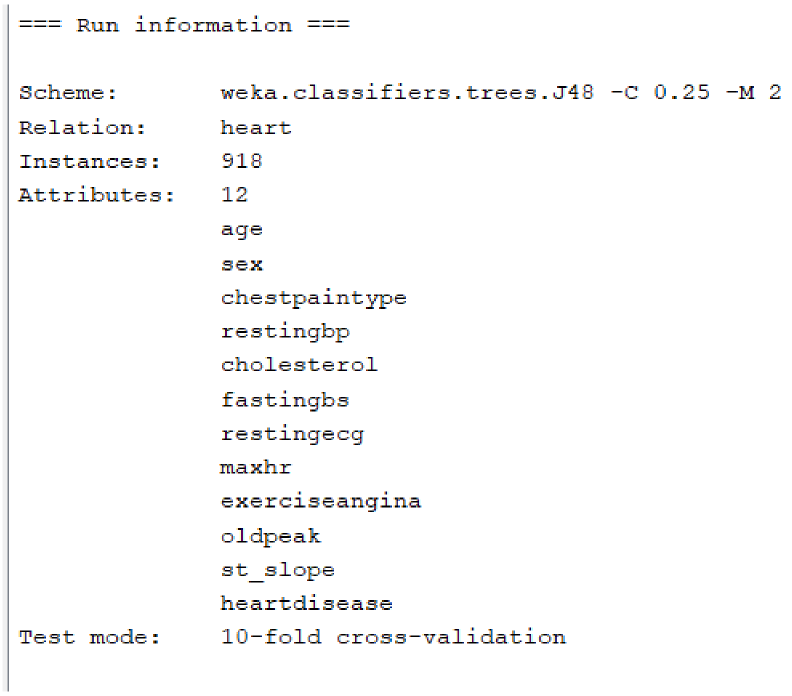


```
=== Summary ===

Correctly Classified Instances         790               86.0566 %
Kappa statistic                          0.7175
Mean absolute error                      0.1575
Root mean squared error                  0.3395
Relative absolute error                 31.8615 %
Root relative squared error             68.2824 %
Total Number of Instances              918

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,837    0,120    0,849      0,837    0,843      0,718    0,918     0,913     0
                 0,880    0,163    0,870      0,880    0,875      0,718    0,918     0,912     1
Weighted Avg.    0,861    0,144    0,860      0,861    0,860      0,718    0,918     0,913

=== Confusion Matrix ===

   a   b   <-- classified as
 343  67 |   a = 0
  61 447 |   b = 1
```

According to the summary analysis, the model has achieved an accuracy of 86.0566% in the test set. From the total 918 instances, 790 of them are correctly classified.

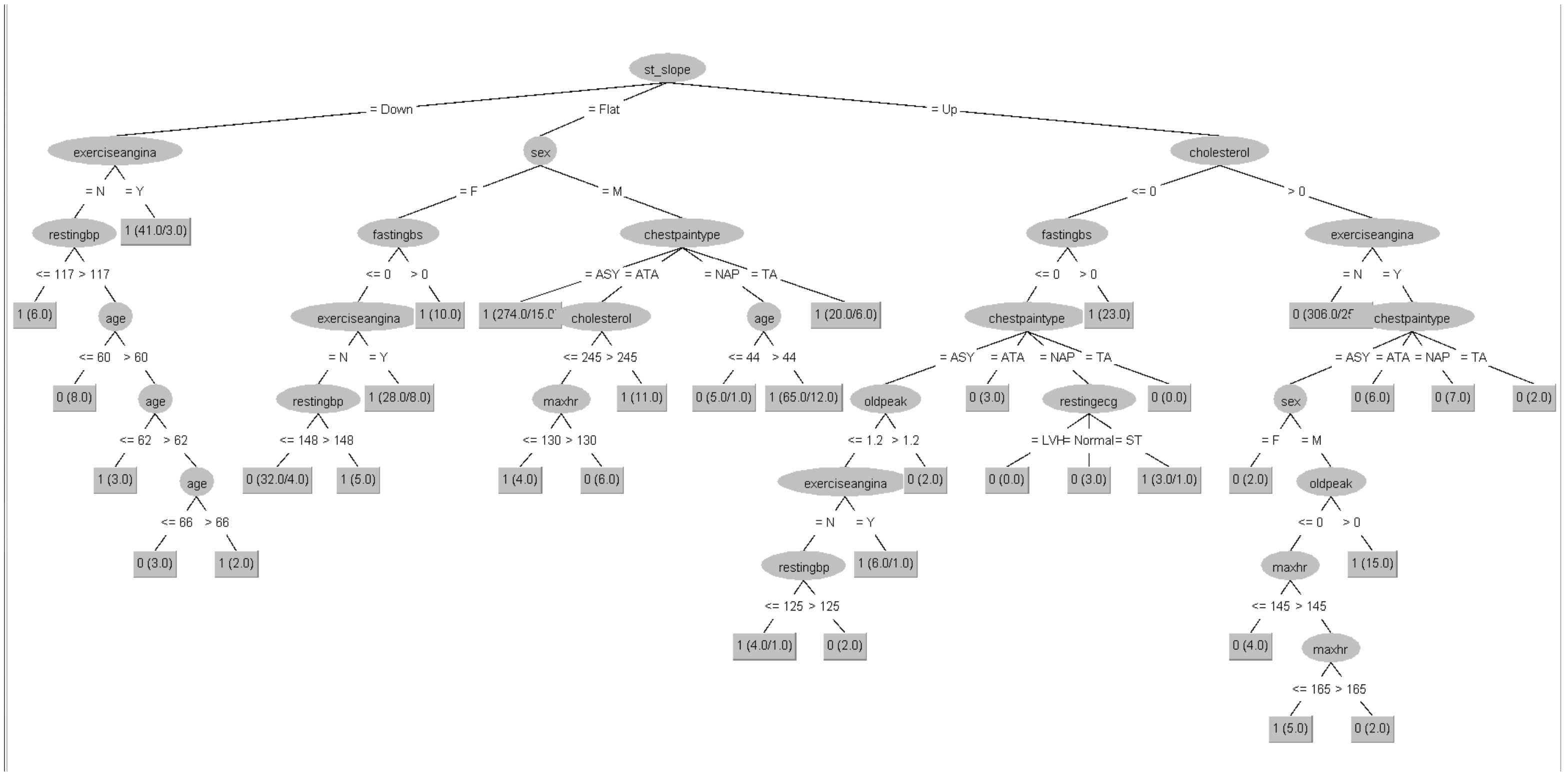
2.2.2 J48 (Decision Tree Classifier)

Run Information:

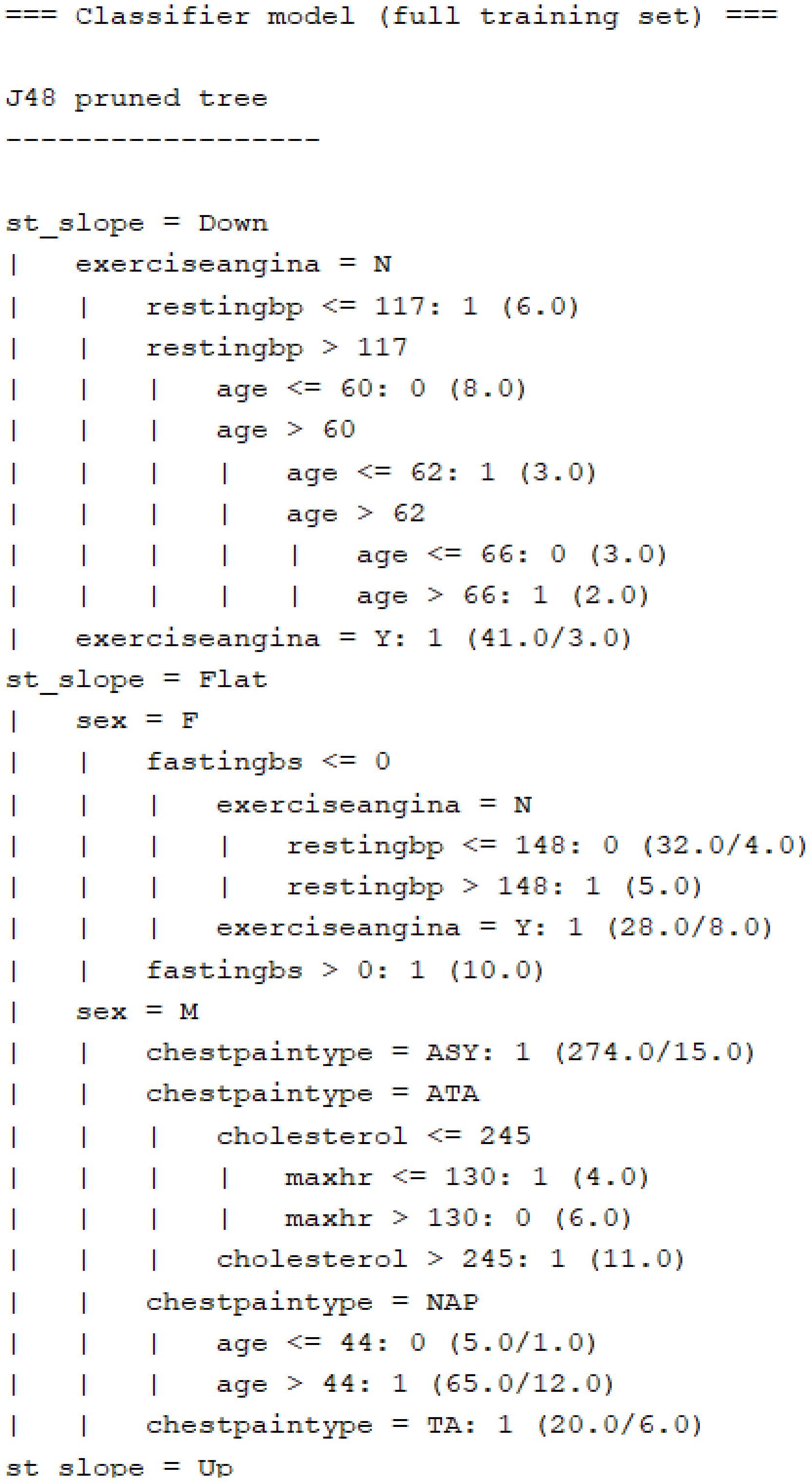


```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     heart
Instances:    918
Attributes:   12
              age
              sex
              chestpaintype
              restingbp
              cholesterol
              fastingbs
              restingecg
              maxhr
              exerciseangina
              oldpeak
              st_slope
              heartdisease
Test mode:    10-fold cross-validation
```

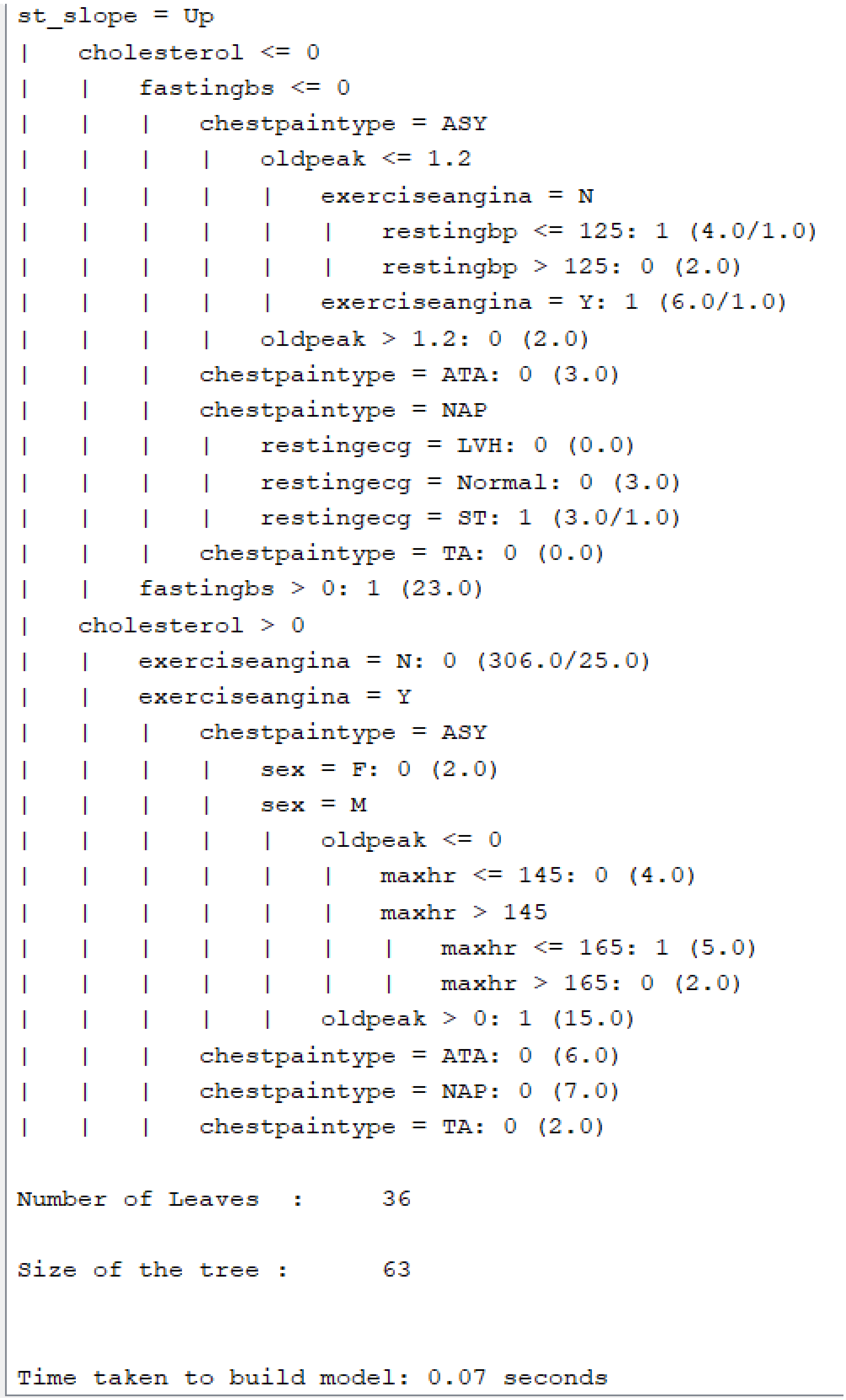
Tree Version:



Training set:

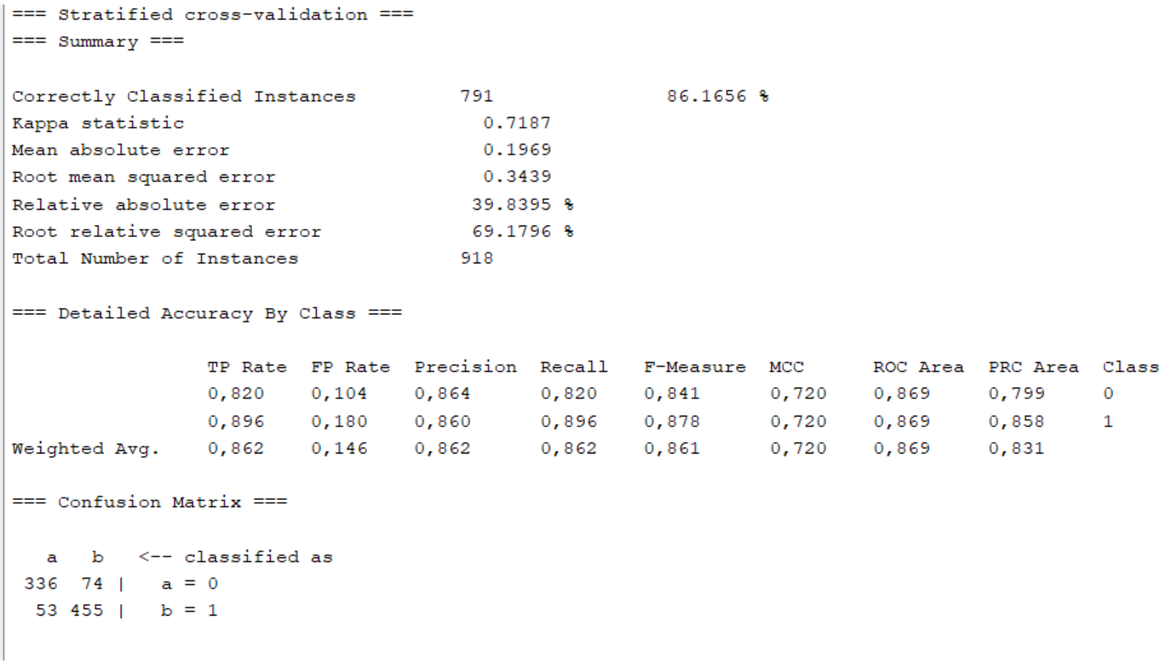


```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

st_slope = Down
|   exerciseangina = N
|   |   restingbp <= 117: 1 (6.0)
|   |   restingbp > 117
|   |   |   age <= 60: 0 (8.0)
|   |   |   age > 60
|   |   |   |   age <= 62: 1 (3.0)
|   |   |   |   age > 62
|   |   |   |   |   age <= 66: 0 (3.0)
|   |   |   |   |   age > 66: 1 (2.0)
|   exerciseangina = Y: 1 (41.0/3.0)
st_slope = Flat
|   sex = F
|   |   fastingbs <= 0
|   |   |   exerciseangina = N
|   |   |   |   restingbp <= 148: 0 (32.0/4.0)
|   |   |   |   restingbp > 148: 1 (5.0)
|   |   |   exerciseangina = Y: 1 (28.0/8.0)
|   |   fastingbs > 0: 1 (10.0)
|   sex = M
|   |   chestpaintype = ASY: 1 (274.0/15.0)
|   |   chestpaintype = ATA
|   |   |   cholesterol <= 245
|   |   |   |   maxhr <= 130: 1 (4.0)
|   |   |   |   maxhr > 130: 0 (6.0)
|   |   |   cholesterol > 245: 1 (11.0)
|   |   chestpaintype = NAP
|   |   |   age <= 44: 0 (5.0/1.0)
|   |   |   age > 44: 1 (65.0/12.0)
|   |   chestpaintype = TA: 1 (20.0/6.0)
st_slope = Up
```



```
st_slope = Up
|   cholesterol <= 0
|   |   fastingbs <= 0
|   |   |   chestpaintype = ASY
|   |   |   |   oldpeak <= 1.2
|   |   |   |   |   exerciseangina = N
|   |   |   |   |   |   restingbp <= 125: 1 (4.0/1.0)
|   |   |   |   |   |   restingbp > 125: 0 (2.0)
|   |   |   |   |   exerciseangina = Y: 1 (6.0/1.0)
|   |   |   |   oldpeak > 1.2: 0 (2.0)
|   |   |   chestpaintype = ATA: 0 (3.0)
|   |   |   chestpaintype = NAP
|   |   |   |   restingecg = LVH: 0 (0.0)
|   |   |   |   restingecg = Normal: 0 (3.0)
|   |   |   |   restingecg = ST: 1 (3.0/1.0)
|   |   |   chestpaintype = TA: 0 (0.0)
|   |   fastingbs > 0: 1 (23.0)
|   cholesterol > 0
|   |   exerciseangina = N: 0 (306.0/25.0)
|   |   exerciseangina = Y
|   |   |   chestpaintype = ASY
|   |   |   |   sex = F: 0 (2.0)
|   |   |   |   sex = M
|   |   |   |   |   oldpeak <= 0
|   |   |   |   |   |   maxhr <= 145: 0 (4.0)
|   |   |   |   |   |   maxhr > 145
|   |   |   |   |   |   |   maxhr <= 165: 1 (5.0)
|   |   |   |   |   |   |   maxhr > 165: 0 (2.0)
|   |   |   |   |   oldpeak > 0: 1 (15.0)
|   |   |   chestpaintype = ATA: 0 (6.0)
|   |   |   chestpaintype = NAP: 0 (7.0)
|   |   |   chestpaintype = TA: 0 (2.0)

Number of Leaves  :     36

Size of the tree :      63


Time taken to build model: 0.07 seconds
```

Summary:

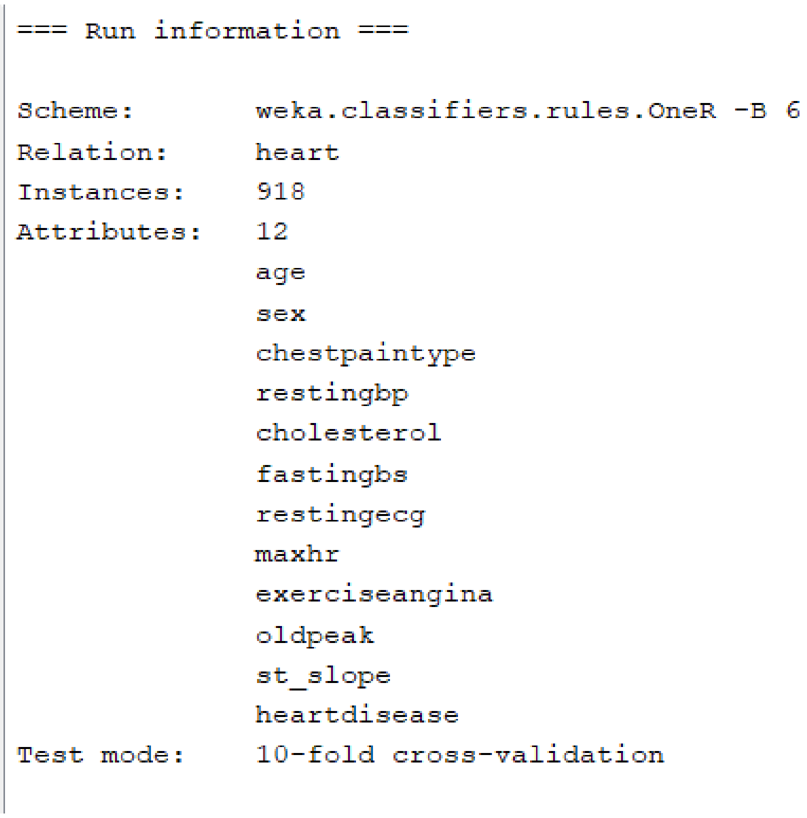


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         791               86.1656 %
Kappa statistic                          0.7187
Mean absolute error                      0.1969
Root mean squared error                  0.3439
Relative absolute error                 39.8395 %
Root relative squared error             69.1796 %
Total Number of Instances              918

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,820    0,104    0,864      0,820    0,841      0,720    0,869     0,799     0
                 0,896    0,180    0,860      0,896    0,878      0,720    0,869     0,858     1
Weighted Avg.    0,862    0,146    0,862      0,862    0,861      0,720    0,869     0,831

=== Confusion Matrix ===

   a   b   <-- classified as
 336  74 |   a = 0
  53 455 |   b = 1
```

According to the summary analysis, the model has achieved an accuracy of 86.1656% in the test set. From the total 918 instances, 791 of them are correctly classified.

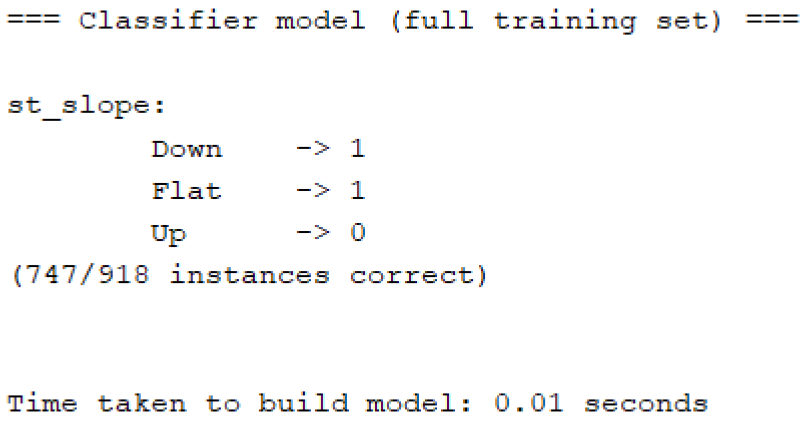
2.2.3 OneR (Rule-based Classifier)

Run Information:



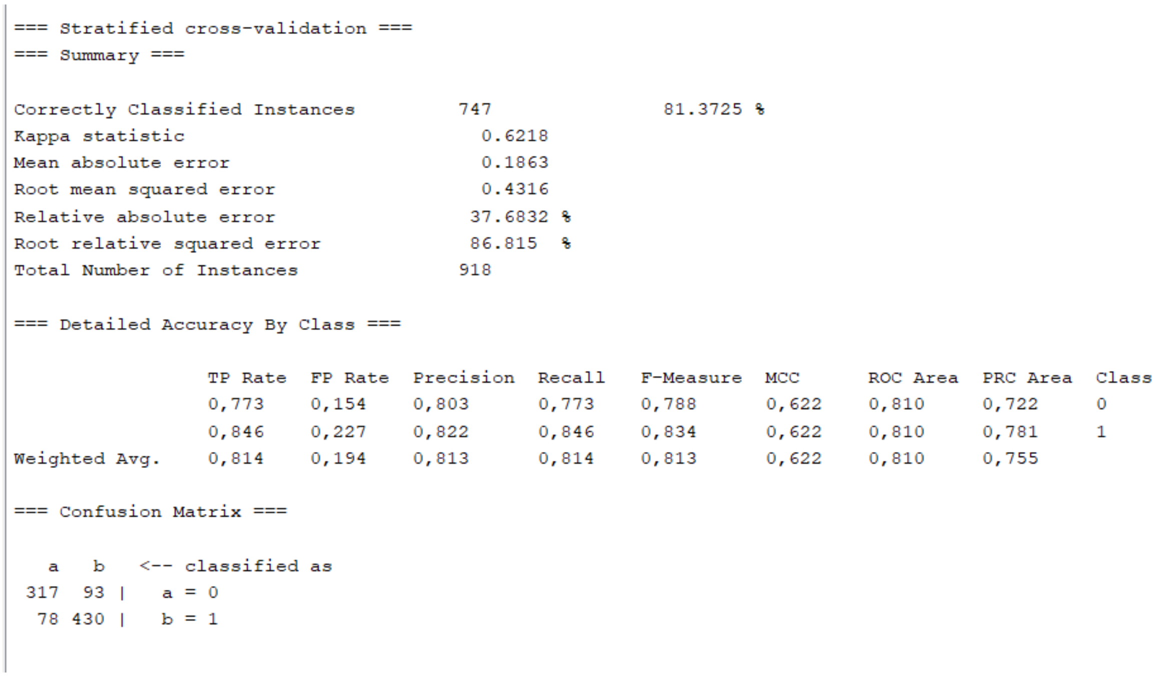
```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     heart
Instances:    918
Attributes:   12
              age
              sex
              chestpaintype
              restingbp
              cholesterol
              fastingbs
              restingecg
              maxhr
              exerciseangina
              oldpeak
              st_slope
              heartdisease
Test mode:    10-fold cross-validation
```

Training set:



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

st_slope:
        Down    -> 1
        Flat    -> 1
        Up      -> 0
(747/918 instances correct)


Time taken to build model: 0.01 seconds
```

Summary:

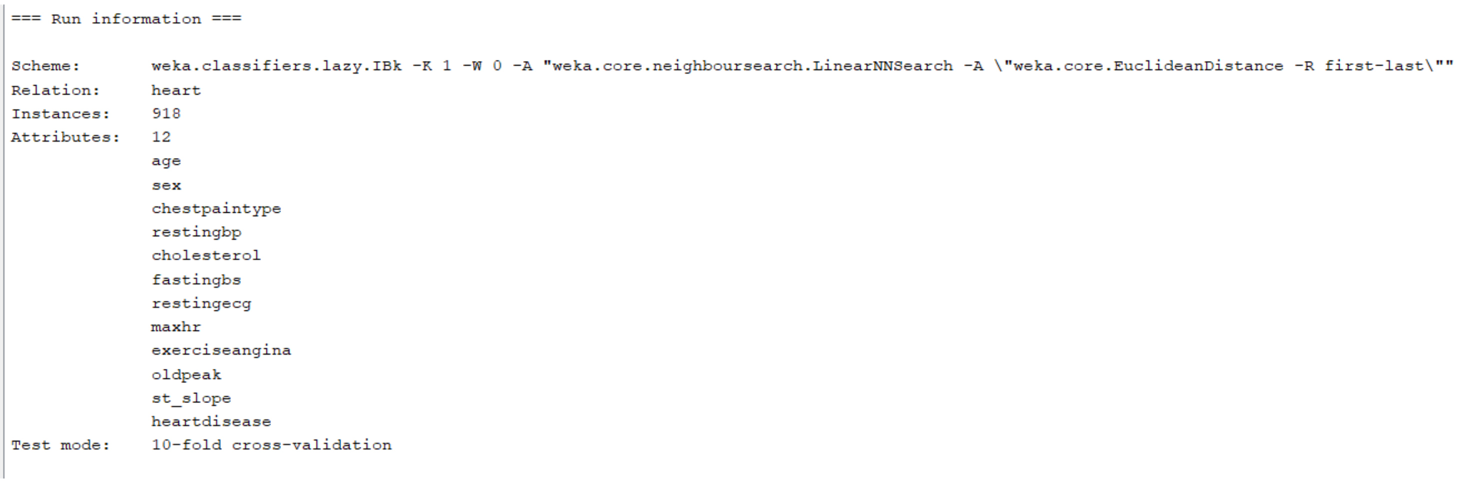


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         747               81.3725 %
Kappa statistic                          0.6218
Mean absolute error                      0.1863
Root mean squared error                  0.4316
Relative absolute error                 37.6832 %
Root relative squared error             86.815  %
Total Number of Instances              918

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,773    0,154    0,803      0,773    0,788      0,622    0,810     0,722     0
                 0,846    0,227    0,822      0,846    0,834      0,622    0,810     0,781     1
Weighted Avg.    0,814    0,194    0,813      0,814    0,813      0,622    0,810     0,755

=== Confusion Matrix ===

   a   b   <-- classified as
 317  93 |   a = 0
  78 430 |   b = 1
```

According to the summary analysis, the model has achieved an accuracy of 81.3725% in the test set. From the total 918 instances, 747 of them are correctly classified.

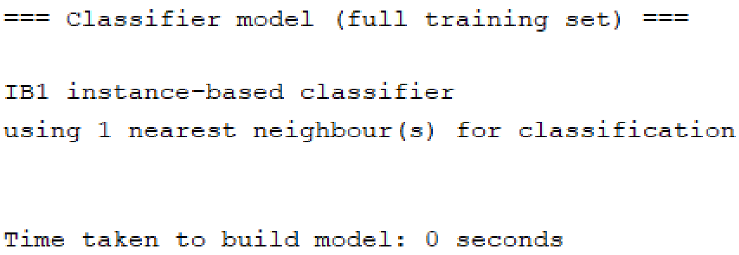
2.2.4 IBk (k-Nearest Neighbor Classifier)

Run Information:



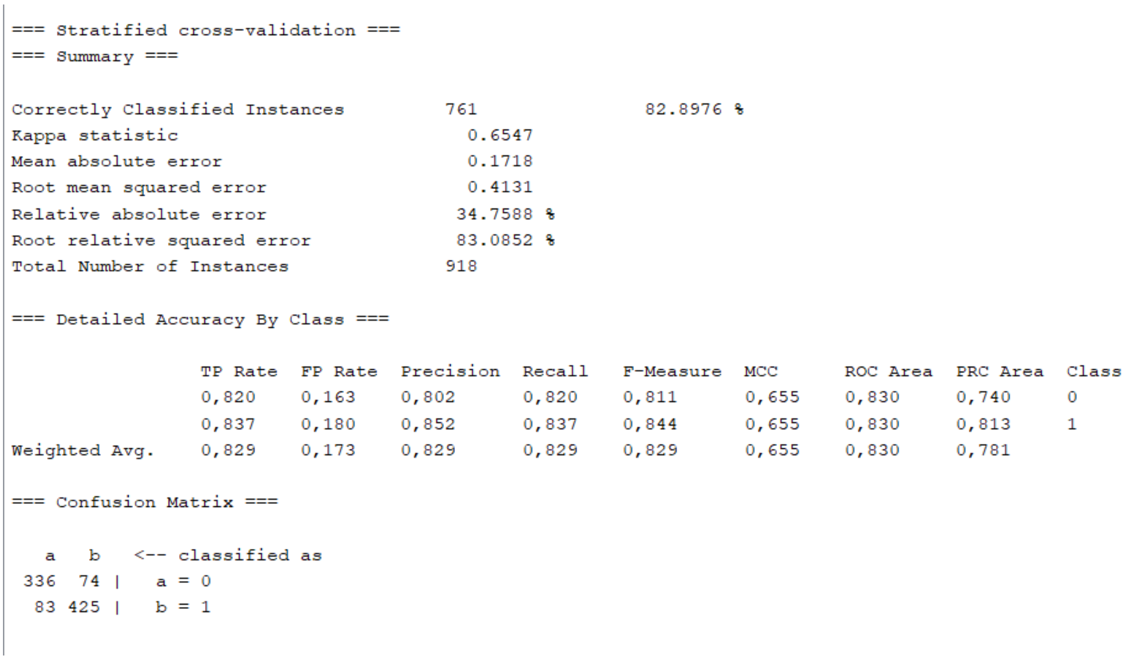
```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     heart
Instances:    918
Attributes:   12
              age
              sex
              chestpaintype
              restingbp
              cholesterol
              fastingbs
              restingecg
              maxhr
              exerciseangina
              oldpeak
              st_slope
              heartdisease
Test mode:    10-fold cross-validation
```

Training set:



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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds
```

Summary:



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

Correctly Classified Instances         761               82.8976 %
Kappa statistic                          0.6547
Mean absolute error                      0.1718
Root mean squared error                  0.4131
Relative absolute error                 34.7588 %
Root relative squared error             83.0852 %
Total Number of Instances              918

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,820    0,163    0,802      0,820    0,811      0,655    0,830     0,740     0
                 0,837    0,180    0,852      0,837    0,844      0,655    0,830     0,813     1
Weighted Avg.    0,829    0,173    0,829      0,829    0,829      0,655    0,830     0,781

=== Confusion Matrix ===

   a   b   <-- classified as
 336  74 |   a = 0
  83 425 |   b = 1
```

According to the summary analysis, the model has achieved an accuracy of 82.8976% in the test set. From the total 918 instances, 761 of them are correctly classified.

3. COMPARISON

3.1 Naive Bayes (Naive Bayes Classifier)

Run Information:

BREAST CANCER	HEART DISEASE
Relations: breast-cancer Instances: 286 Attributes: 10	Relations: heart Instances: 918 Attributes: 12

Training set:

BREAST CANCER	HEART DISEASE
Resulted in: 0 seconds	Resulted in: 0.03 seconds

Summary:

BREAST CANCER	HEART DISEASE
71.6783 % Correctly Classified Instances 205 Kappa statistic 0.2857 Mean absolute error 0.3272 Root mean squared error 0.4534 Relative absolute error 78.2086 % Root relative squared error 99.1872 % Total Number of Instances 286	**86.0566 %** Correctly Classified Instances 790 Kappa statistic 0.7175 Mean absolute error 0.1575 Root mean squared error 0.3395 Relative absolute error 31.8615 % Root relative squared error 68.2824 % Total Number of Instances 918

3.2 J48 (Decision Tree Classifier)

Run Information:

BREAST CANCER	HEART DISEASE
Relations: breast-cancer Instances: 286 Attributes: 10	Relations: heart Instances: 918 Attributes: 12

Training set:

BREAST CANCER	HEART DISEASE
Resulted in: 0,02 seconds	Resulted in: 0.07 seconds

Summary:

BREAST CANCER	HEART DISEASE
75.5245 %	**86.1656 %**
Correctly Classified Instances 216	Correctly Classified Instances 791
Kappa statistic 0.2826	Kappa statistic 0.7187
Mean absolute error 0.3676	Mean absolute error 0.1969
Root mean squared error 0.4324	Root mean squared error 0.3439
Relative absolute error 87.8635 %	Relative absolute error 39.8395 %
Root relative squared error 94.6093 %	Root relative squared error 69.1796 %
Total Number of Instances 286	Total Number of Instances 918

3.3 OneR (Rule-based Classifier)

Run Information:

BREAST CANCER	HEART DISEASE
Relations: breast-cancer	Relations: heart
Instances: 286	Instances: 918
Attributes: 10	Attributes: 12

Training set:

BREAST CANCER	HEART DISEASE
Resulted in: 0 seconds	Resulted in: 0.01 seconds

Summary:

BREAST CANCER	HEART DISEASE
65.7343 %	**81.3725 %**
Correctly Classified Instances 188	Correctly Classified Instances 747
Kappa statistic 0.0936	Kappa statistic 0.6218
Mean absolute error 0.3427	Mean absolute error 0.1863
Root mean squared error 0.5854	Root mean squared error 0.4316
Relative absolute error 81.8943 %	Relative absolute error 37.6832 %
Root relative squared error 128.0681 %	Root relative squared error 86.815 %
Total Number of Instances 286	Total Number of Instances 918

3.4 IBk (k-Nearest Neighbor Classifier)

Run Information:

BREAST CANCER	HEART DISEASE
Relations: breast-cancer Instances: 286 Attributes: 10	Relations: heart Instances: 918 Attributes: 12

Training set:

BREAST CANCER	HEART DISEASE
Resulted in: 0 seconds	Resulted in: 0 seconds

Summary:

BREAST CANCER	HEART DISEASE
72.3776% Correctly Classified Instance 207 Kappa statistic 0.2438 Mean absolute error 0.3257 Root mean squared error 0.5101 Relative absolute error 77.8513 % Root relative squared error 111.6114 % Total Number of Instances 286	**82.8976 %** Correctly Classified Instances 761 Kappa statistic 0.6547 Mean absolute error 0.1718 Root mean squared error 0.4131 Relative absolute error 34.7588 % Root relative squared error 83.0852 % Total Number of Instances 918

4. CONCLUSION

The results of the datasets have been analyzed and models are compared with each other. In this study "breast-cancer" data set and "heart" data set used in WEKA Data Mining application for observing the classification results. Their attributes are close to each other. (10 attributes in breast-cancer dataset and 12 attributes in heart data set.

Four classification methods used (Bayes-Naive Bayes, Tree-.J48, Rules-OneR and Lazy-IBk).

The J48 gave the most accurate results for both datasets. However, the second accurate one was different. Naïve Bayes was better for second best accuracy with datasets have more instances (Heart Disease dataset in this study.) As for this, IBk was better for second best accuracy with datasets have less instances. (Breast Cancer dataset in this study.)

Consequently, large datasets give better results according to the smaller ones.

5. REFERENCES

[1]: https://github.com/renatopp/arff-datasets/blob/master/classification/breast.cancer.arff

[2]: https://www.kaggle.com/fedesoriano/heart-failure-prediction