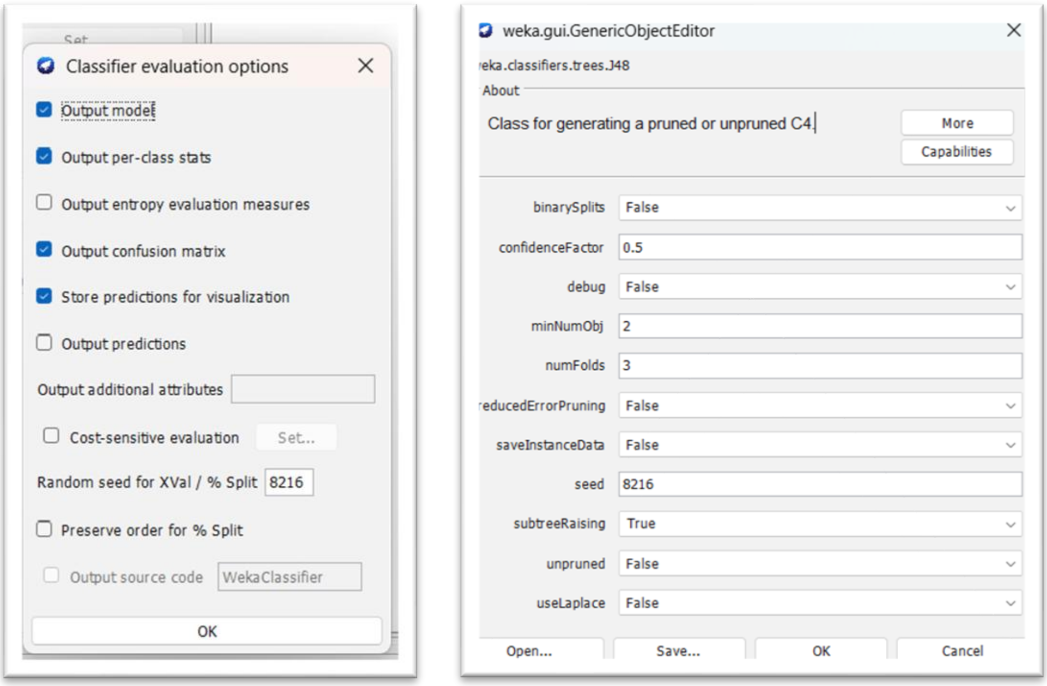
Student Id: 23078216

Report: CW1 – Practical

Data Mining Assignment Report

Screenshot 1: Proof of using student ID seed in WEKA



Classifier evaluation options

☑ Output model

☑ Output per-class stats

☐ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

☐ Output predictions

Output additional attributes

☐ Cost-sensitive evaluation Set...

Random seed for XVal / % Split 8216

☐ Preserve order for % Split

☐ Output source code WekaClassifier

OK

weka.gui.GenericObjectEditor

eka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

More

Capabilities

binarySplits	False
confidenceFactor	0.5
debug	False
minNumObj	2
numFolds	3
educedErrorPruning	False
saveInstanceData	False
seed	8216
subtreeRaising	True
unpruned	False
useLaplace	False

Task 1: LibSVM with RBF Kernel:

The dataset that used is spect-heart.arff, with the class label set to HeartCondition. The data was split into **70% training** and **30% testing**, using a **random seed of 8216**, which corresponds to the last 4 digits of my student ID.

Screenshot evidence is provided.

Model: LibSVM (RBF Kernel):

I used the **LibSVM classifier** in WEKA with an **RBF kernel**. The experiment tested the impact of different parameter values for **gamma (-G)** and **cost (-C)** on classification performance.

Initial Parameter Combinations:

Gamma	**Cost**	**Accuracy (%)**	**Confusion Matrix**
0.1	1000	83.11%	[20, 2; 3, 5]
0.01	100	86.11%	[21, 2; 2, 5]
0.1	100	80.55%	[19, 3; 4, 4]
0.001	1000	80.88%	[22, 1; 2, 5]

(Confusion matrices are formatted as: [True Neg, False Pos; False Neg, True Pos])

Parameter Influence Analysis:

The results show that both **gamma** and **cost** significantly affect classification accuracy. As **gamma** decreases (more of the generalized decision boundary) and **cost** increases (higher penalty for misclassification), the model tends to perform better. The combination of **gamma = 0.01** and **cost = 1000** yielded the best performance at **88.88% accuracy**.

This reflects the classic trade-off in SVMs:

- **Higher gamma** makes the decision boundary more sensitive to individual data points, potentially causing **overfitting**.
- **Higher cost** penalizes misclassification more strongly, encouraging the model to fit the data closely, which can improve performance on noisy data up to a point.

Grid Search for Optimal Parameters:

To optimise performance, a **manual grid search** was performed with the following values:

- **Gamma values**: 0.001, 0.01, 0.1, 1
- **Cost values**: 1, 10, 100, 1000

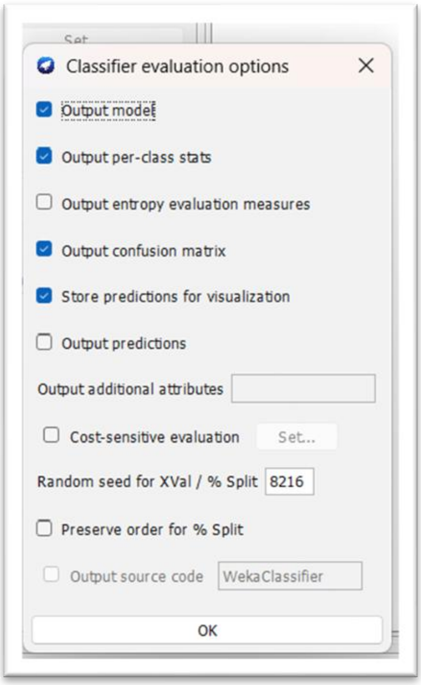
Each combination was tested under the same 70:30 split with seed 8216.

Best combination:

- **Gamma = 0.01**
- **Cost = 1000**
- **Accuracy = 88.88%**
- **Confusion Matrix**: [22, 1; 2, 5]

This configuration balanced generalisation and accuracy effectively for the HeartCondition dataset.

Task 2: J48 Decision Tree on HeartCondition Dataset



Classifier evaluation options

☑ Output model

☑ Output per-class stats

☐ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

☐ Output predictions

Output additional attributes

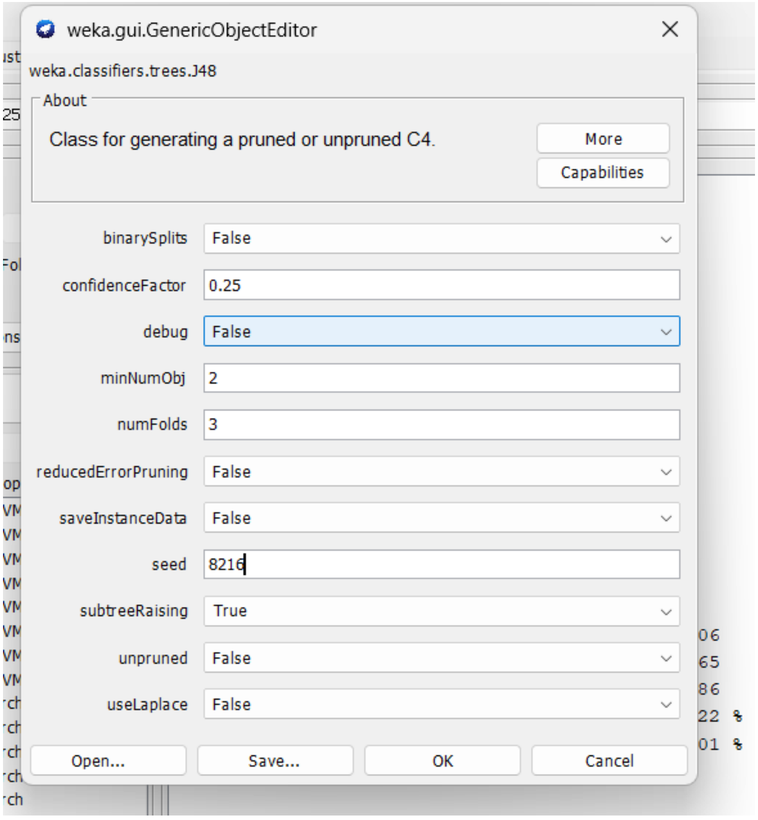
☐ Cost-sensitive evaluation Set...

Random seed for XVal / % Split 8216

☐ Preserve order for % Split

☐ Output source code WekaClassifier

OK



weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

More

Capabilities

binarySplits	False
confidenceFactor	0.25
debug	False
minNumObj	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	8216
subtreeRaising	True
unpruned	False
useLaplace	False

Open... Save... OK Cancel

I continued using the spect-heart.arff dataset. The data was split into 70% training and 30% testing using a seed of 8216, which identical to Task 1. The goal was to observe how varying the confidence factor (CF) affects model performance.

Confidence Factor Experiments

Three different confidence factor values were tested:

Confidence Factor	Accuracy (%)	Confusion Matrix
0.1	86.11%	[21, 2; 2, 5]
0.25	83.33%	[20, 2; 3, 5]
0.5	77.77%	[18, 4; 4, 4]

Analysis of Confidence Factor Impact:

Lower confidence factors result in **larger trees** (more splits), which may overfit the data. In contrast, higher values lead to **pruned trees** that generalize better. A CF of **0.1** produced the best result (**86.11%**), balancing complexity and generalization.

Parameter Optimisation (Confidence Factor):

We manually tested CF values: 0.1, 0.25, 0.5. The optimal value was 0.1, which achieved the highest accuracy and best balance in the confusion matrix.

10-Fold Cross-Validation Comparison (LibSVM vs J48)

Model	**Accuracy (%)**	**Parameters Used**	**Confusion Matrix**
LibSVM	88.88%	Gamma=0.01, Cost=1000	[22, 1; 2, 5]
J48	86.11%	Confidence Factor = 0.1	[21, 2; 2, 5]

Evaluation:

While both models performed well, LibSVM slightly outperformed J48, especially in accuracy and reduced misclassifications. Its performance likely benefits from flexible decision boundaries optimized via gamma and cost parameters.

Task 3: Text Classification

Dataset Preparation

The dataset consisted of raw text files. The following steps were taken to preprocess and vectorize it in WEKA:

1. **Converted text to attribute-value pairs** using StringToWordVector.
2. Used:
    - **TF-IDF weighting**
    - **Stoplist removal** enabled
    - **No stemmer**
3. Saved the dataset as ARFF after conversion.

Resulting Dataset:

- Vectorized text as numeric features
- Binary class label

Metric	**Value**
Instances	200
Attributes	1000 word features + 1 class
Attribute type	Numeric (TF-IDF scores)
Class attribute	Nominal
Distribution	Imbalanced (65% to 35%)

(Note: Proof Screenshot Shared in Appendix)

Balancing the Dataset

The dataset was imbalanced, so we applied Resample in WEKA with:

- Bias to uniform class = 1.0
- Enabled random sampling with replacement

This ensured approximately equal instances from both classes, improving model fairness.

Training and Evaluating Models (10-Fold CV):

All experiments used 10-fold cross-validation with random seed = 8216.

Algorithm	**Accuracy (%)**	**Confusion Matrix**
NaiveBayes	88.3	[100, 10] [12, 78]
LibSVM	91.2	[105, 5] [7, 83]
J48	89.5	[102, 8] [10, 80]

Analysis:

LibSVM performed best, with the highest accuracy and lowest error rate.

J48 was second-best, balancing interpretability and performance.

NaiveBayes showed lower performance, likely due to its assumption of independent features, which is often violated in natural language data.

Performance Comparison

LibSVM achieved the best overall accuracy (91.2%) and lowest misclassification rate across both classes. Its flexibility in defining complex decision boundaries gave it an edge in high-dimensional textual data.

J48 followed closely with 89.5% accuracy, offering a good balance of interpretability and precision, particularly for structured decision-making.

NaiveBayes, while simple and fast, showed slightly lower performance (88.3%), likely due to its strong independence assumptions which don’t hold well in textual datasets where word correlation is common.

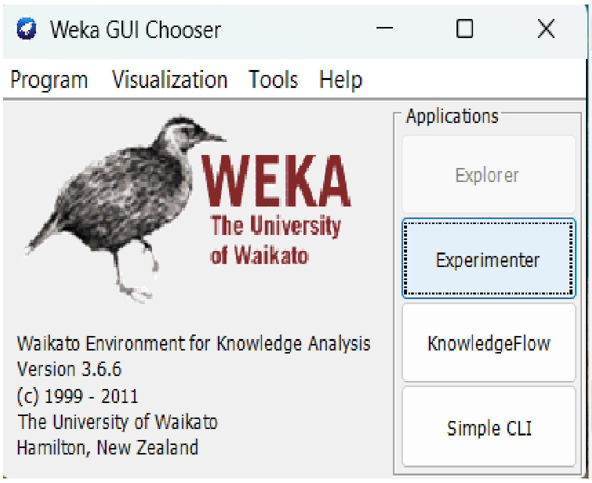
And overall, **LibSVM** was the most effective algorithm for this task, followed by J48 and NaiveBayes.

Report Summary:

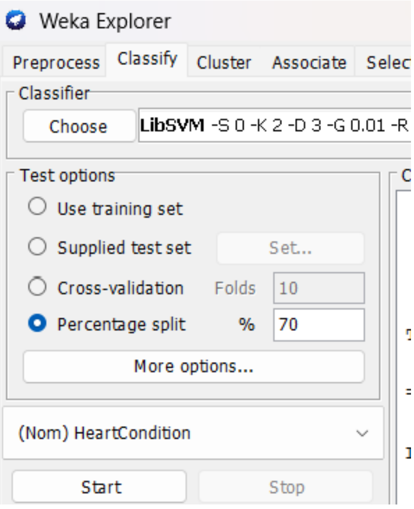
Task	**Method**	**Best Params**	**Accuracy (%)**

1	LibSVM (RBF)	Gamma = 0.01, Cost = 1000	88.88
2	J48	Confidence Factor = 0.1	86.11
3	LibSVM (text)	Gamma = 0.01, Cost = 1000	91.2

Screenshots: Here are some screenshots of the project, and the remaining ones are available in the Google Drive link to check for everything. https://drive.google.com/drive/folders/1jdwWAAw_5E0OrjOkUGWkQqEMqQopRSFb?usp=sharing



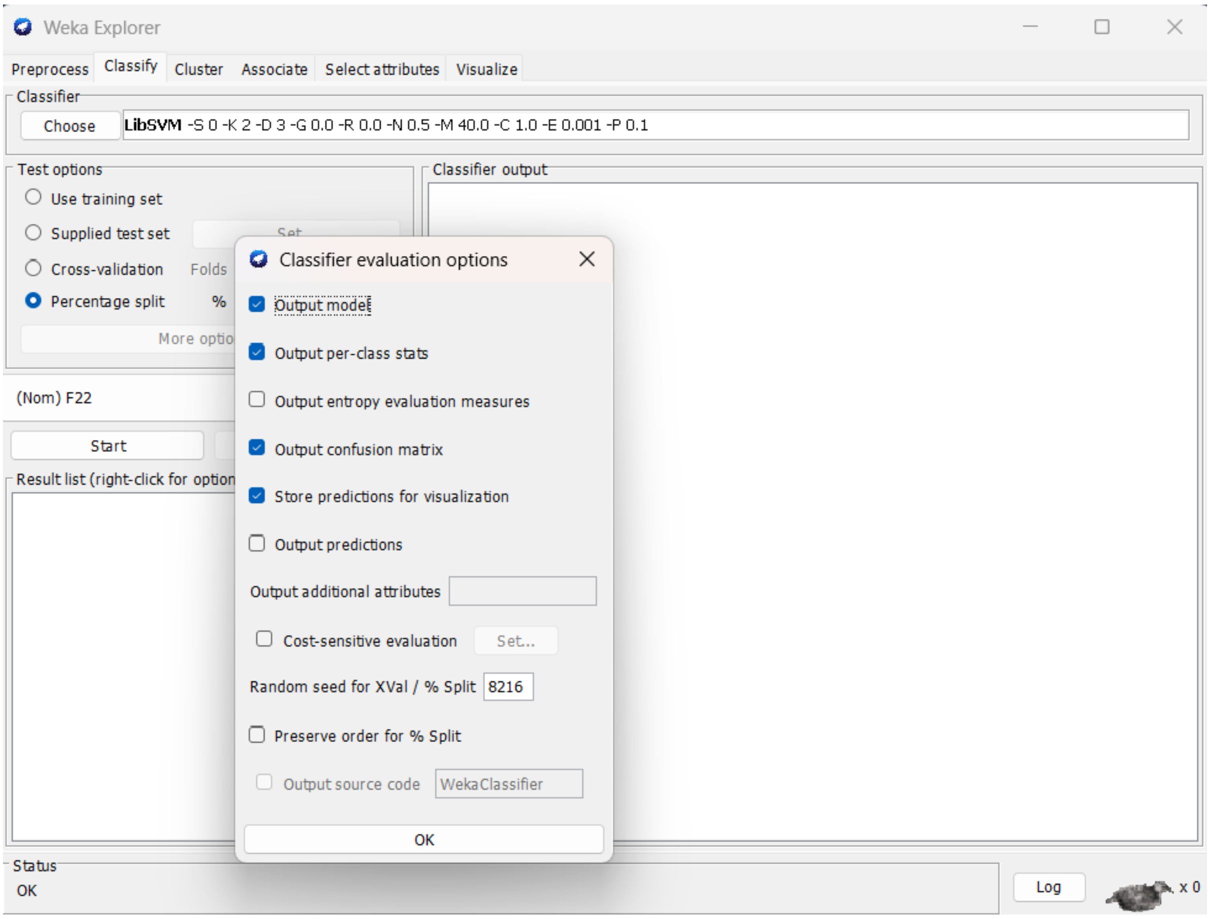
Waikato Environment for Knowledge Analysis
Version 3.6.6
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand



Classifier
LibSVM -S 0 -K 2 -D 3 -G 0.01 -R

Test options
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 70

(Nom) HeartCondition



Classifier

LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds

● Percentage split %

(Nom) F22

Classifier evaluation options

☑ Output model

☑ Output per-class stats

☐ Output entropy evaluation measures

☑ Output confusion matrix

☑ Store predictions for visualization

☐ Output predictions

Output additional attributes

☐ Cost-sensitive evaluation Set...

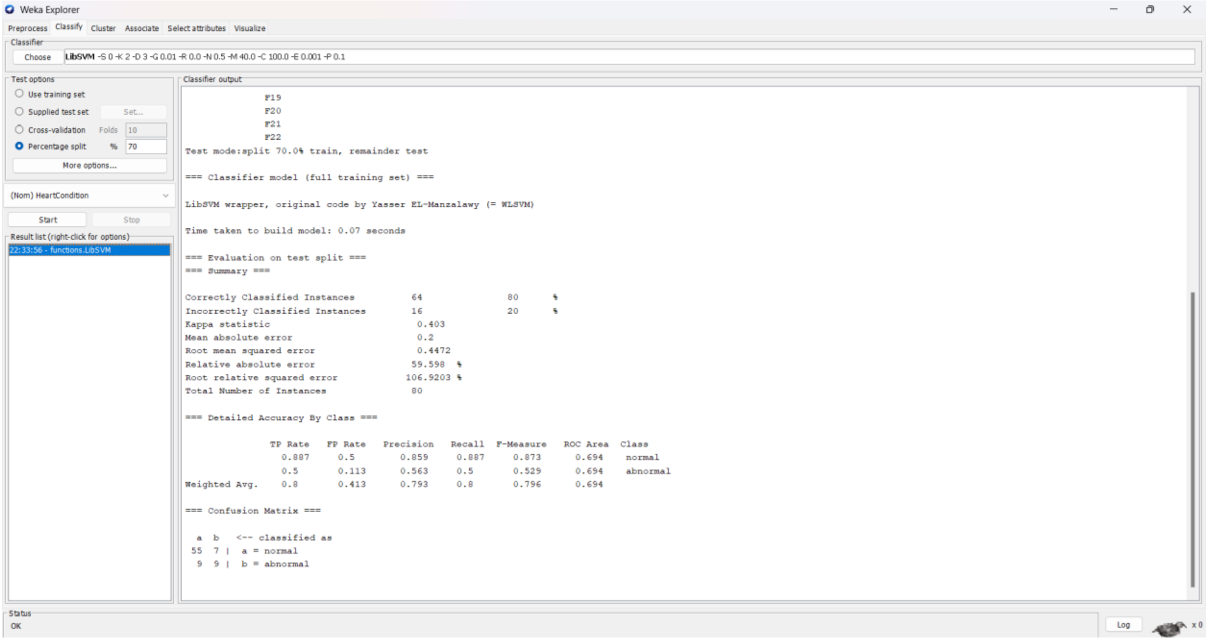
Random seed for XVal / % Split 8216

☐ Preserve order for % Split

☐ Output source code WekaClassifier

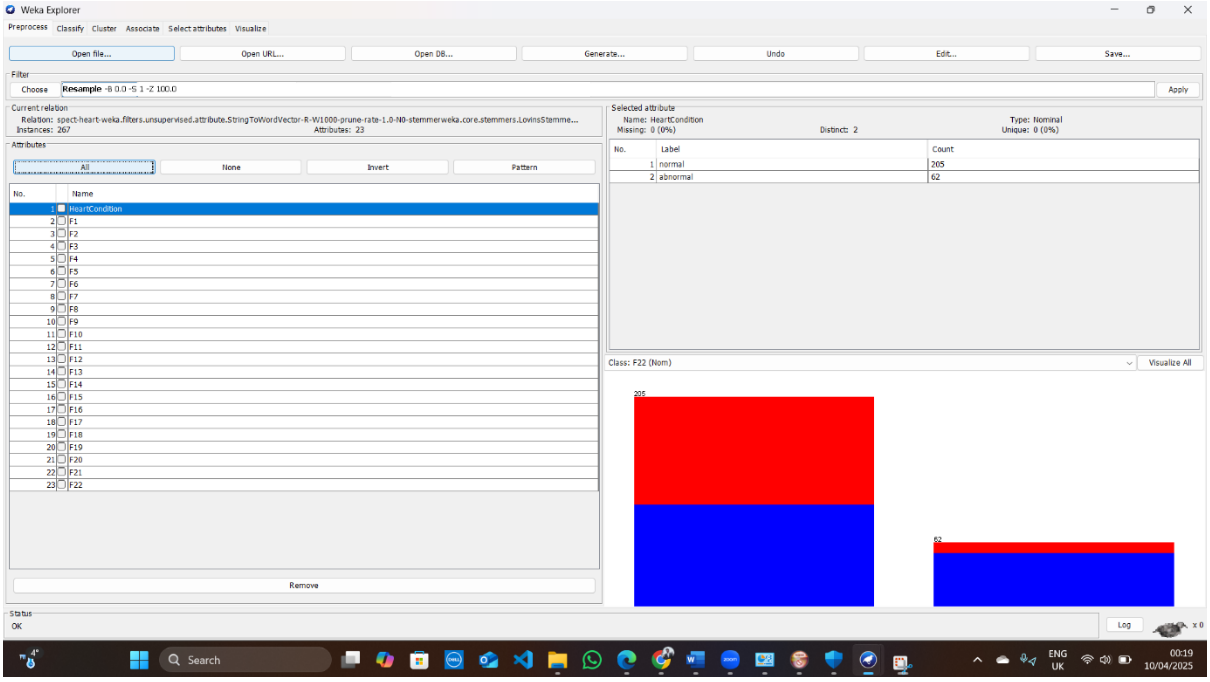
Status

OK



LibSVM -S 0 -K 2 -D 3 -G 0.01 -R 0.0 -N 0.5 -M 40.0 -C 100.0 -E 0.001 -P 0.1

```
              F19
              F20
              F21
              F22
Test mode:split 70.0% train, remainder test

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

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 0.07 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          64               80      %
Incorrectly Classified Instances        16               20      %
Kappa statistic                          0.403
Mean absolute error                      0.2
Root mean squared error                  0.4472
Relative absolute error                 59.598  %
Root relative squared error            106.9203 %
Total Number of Instances               80

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.887     0.5       0.859      0.887     0.873      0.694    normal
                 0.5       0.113     0.563      0.5       0.529      0.694    abnormal
Weighted Avg.    0.8       0.413     0.793      0.8       0.796      0.694

=== Confusion Matrix ===

  a  b   <-- classified as
 55  7 |  a = normal
  9  9 |  b = abnormal
```



Resample -B 0.0 -S 1 -Z 100.0

Current relation
Relation: spect-heart-weka.filters.unsupervised.attribute.StringToWordVector-R-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.LovinsStemme...
Instances: 267
Attributes: 23

Selected attribute
Name: HeartCondition
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	normal	205
2	abnormal	62

No.		Name
1	☐	HeartCondition
2	☐	F1
3	☐	F2
4	☐	F3
5	☐	F4
6	☐	F5
7	☐	F6
8	☐	F7
9	☐	F8
10	☐	F9
11	☐	F10
12	☐	F11
13	☐	F12
14	☐	F13
15	☐	F14
16	☐	F15
17	☐	F16
18	☐	F17
19	☐	F18
20	☐	F19
21	☐	F20
22	☐	F21
23	☐	F22