

FINAL PROJECT REPORT

CS699 - DATA MINING

BY PREETI CHOUBEY

TABLE OF CONTENTS

Dataset.....	3
Data Preprocessing:.....	3
Handling missing values.....	3
Standardization.....	3
Removing Duplicates	3
Finding Outliers.....	3
Remove Outliers	3
Randomization.....	3
Splitting the dataset	3
Feature selection	4
Selected attribute based on CorrelationAttributeEval	5
Selected attribute based on ChiSquaredAttributeEval.....	5
Selected attribute based on InfoGainAttributeEval.....	6
Selected attribute based on GainRatioAttributeEval	6
Selected attribute based on OneRAttributeEval	7
Training.....	7
Classifiers Used	7
Test results of all 25 models.....	8
Correlation attribute selection with Naive Bayes algorithm	8
Correlation attribute selection with J48 Decision Tree algorithm.....	9
Correlation attribute selection with Logistic algorithms.	9
Correlation attribute selection with Support Vector Machine algorithm.	10
Correlation attribute selection with Bagging (Random Forest) algorithm.	10
Chi-Squared Attribute selection with Naive Bayes algorithm:	11
Chi-Squared Attribute selection with Logistic algorithm.....	11
Chi-Squared Attribute selection with Support Vector Machine algorithm	13
Chi-Squared Attribute selection with J48 Decision Tree algorithm	13

Chi-Squared Attribute selection with Bagging (Random Forest) algorithm	14
Info Gain attribute selection with Naive Bayes algorithm.....	14
Info Gain attribute selection with Logistic algorithm.	15
Info Gain attribute selection with J48 Decision Tree algorithm.	15
Info Gain attribute selection with Support Vector Machine algorithm.....	16
Info Gain attribute selection with Bagging (Random Forest) algorithm.....	16
Gain Ratio Attribute selection with Naive Bayes algorithm:	17
Gain Ratio Attribute selection with Logistic algorithm:.....	17
Gain Ratio Attribute selection with J48 Decision Tree algorithm:.....	18
Gain Ratio Attribute selection with Support Vector Machine algorithm:	18
Gain Ratio Attribute selection with Bagging (Random Forest) algorithm:	19
One R attribute selection with Naive Bayes algorithm.....	19
One R attribute selection with Logistic algorithm.	20
One R attribute selection with J48 Decision Tree algorithm.	20
One R attribute selection with Support Vector Machine algorithm.....	21
One R attribute selection with Bagging (Random Forest) algorithm.....	21
Conclusion.....	23

DATASET:

The project-2018-BRFSS-arthritis.csv dataset included 11933 tuples and 108 attributes. Each tuple is a person who participated in the survey and each attribute represents an answer to a survey question. The class attribute is havarth3 and its value is either 1 or 2. 1 means that the individual was diagnosed with arthritis at some point, while 2 means that they haven't.

DATA PREPROCESSING:

HANDLING MISSING VALUES

There are no missing values for havarth3 (the label). For other attributes, I used Weka's ReplaceMissingValues filter. It replaces all missing values for the numeric and nominal attributes in a dataset with the mean and mode from the data, respectively. For this, I followed these steps.

From choose button > weka > Filters > unsupervised > attribute > replaceMissingValues > apply > Save.

STANDARDIZATION

Standardized numeric attributes to zero mean and unit variance.

REMOVING DUPLICATES

Unnecessary, as there are no duplicate entries.

FINDING OUTLIERS

From choose button > WEKA > Filters > unsupervised > attribute > interquartile range > apply.

REMOVE OUTLIERS

choose button > weka > filter > unsupervised > instance > remove with value > click on the filter field (to add outlier attribute indices and nominal indices, I specified it as last value) > apply > remove outliers' attribute.

RANDOMIZATION

It randomly shuffles the order of the instances passed through it.

SPLITTING THE DATASET

I split the data into a training set (66% of the data) and a test set (34% of the data).

To do so, I used the RemovePercentage filter (Weka>filter>unsupervised > instance)

For the training set:

1. Load the full dataset
2. Select the RemovePercentage filter in the preprocess panel

3. Set the split to 34%
4. Apply the filter
5. Save the generated data as project-training.arff

For test set:

1. Load the full dataset
2. Select the RemovePercentage filter
3. Set the invert Selection property to true
4. Apply the filter
5. Save the generated data as project-test.arff

FEATURE SELECTION

For attribute selection, I used the following attribute selection methods:

- CorrelationAttributeEval: Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class.
- ChiSquaredAttributeEval: Evaluates the worth of an attribute by computing the value of the chi-squared statistic for the class.
- InfoGainAttributeEval: Evaluates the worth of an attribute by measuring the information gain for the class.
- GainRatioAttributeEval: Evaluates the worth of an attribute by measuring the gain ratio for the class
- OneRAttributeEval: Evaluates the worth of an attribute by using the OneR classifier.

SELECTED ATTRIBUTE BASED ON CORRELATIONATTRIBUTEVAL

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AttributeSelectedClassifier -E "weka.attributeSelection.CorrelationAttributeEval" -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 23" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

- 15:37:32 - meta.AttributeSelectedClassifier
- 15:37:46 - meta.AttributeSelectedClassifier
- 15:38:00 - meta.AttributeSelectedClassifier
- 15:38:13 - meta.AttributeSelectedClassifier
- 15:38:27 - meta.AttributeSelectedClassifier
- 15:38:40 - meta.AttributeSelectedClassifier
- 15:39:45 - meta.AttributeSelectedClassifier
- 15:39:57 - meta.AttributeSelectedClassifier
- 15:40:10 - meta.AttributeSelectedClassifier
- 15:40:23 - meta.AttributeSelectedClassifier

Classifier output

Attribute Evaluator (supervised, Class (nominal): 108 havarth3):
Correlation Ranking Filter

Ranked attributes:

0.352	22 diffwalk
0.351	64 x.age80
0.348	66 x.ageg5yr
0.309	2 employl
0.275	87 x.rfhlth
0.268	67 x.age65yr
0.24	97 x.hcvu651
0.202	20 pneuvac4
0.201	102 x.exteth3
0.195	3 income2
0.193	95 x.phys14d
0.186	25 diffalon
0.185	31 phsylth
0.178	6 children
0.167	46 chccopd1
0.166	24 diffdres
0.163	62 x.age.g
0.162	29 diabete3
0.157	69 x.chldcnt
0.154	36 chckdny1
0.151	104 x.michd
0.15	27 rmvteth4
0.147	43 checkup1

Selected attributes: 22,64,66,2,87,67,97,20,102,3,95,25,31,6,46,24,62,29,69,36,104,27,43 : 23

SELECTED ATTRIBUTE BASED ON CHISQUAREDATTRIBUTEVAL

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AttributeSelectedClassifier -E "weka.attributeSelection.ChiSquaredAttributeEval" -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 25" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

- 09:18:07 - meta.AttributeSelectedClassifier
- 09:18:20 - meta.AttributeSelectedClassifier
- 09:18:58 - meta.AttributeSelectedClassifier
- 09:19:13 - meta.AttributeSelectedClassifier
- 09:19:25 - meta.AttributeSelectedClassifier
- 09:19:41 - meta.AttributeSelectedClassifier
- 09:20:16 - meta.AttributeSelectedClassifier
- 09:20:30 - meta.AttributeSelectedClassifier

Classifier output

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 108 havarth3):
Chi-squared Ranking Filter

Ranked attributes:

305.6896	22 diffwalk
296.6604	64 x.age80
294.821	62 x.age.g
293.4988	66 x.ageg5yr
274.7443	2 employl
274.3673	34 genhlth
221.5711	31 phsylth
206.8342	95 x.phys14d
185.8565	87 x.rfhlth
177.9911	97 x.hcvu651
177.056	67 x.age65yr
146.5377	20 pneuvac4
124.9581	27 rmvteth4
107.0976	102 x.exteth3
102.3139	11 marital
96.5154	71 x.incomg
86.8385	3 income2
86.2643	25 diffalon
82.4374	69 x.chldcnt
78.304	6 children

Selected attributes: 22,64,62,66,2,34,31,95,87,97,67,20,27,102,11,71,3,25,69,6 : 20

SELECTED ATTRIBUTE BASED ON INFOGAINATTRIBUTEVAL

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AttributeSelectedClassifier -E "weka.attributeSelection.InfoGainAttributeEval" -S "weka.attributeSelection.Ranker" -T -1.7976931348623157E308 -N 23" -W weka.classifier

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

16:36:24 - meta.AttributeSelectedClassifier
16:36:37 - meta.AttributeSelectedClassifier
16:36:49 - meta.AttributeSelectedClassifier
16:37:01 - meta.AttributeSelectedClassifier
16:37:14 - meta.AttributeSelectedClassifier
16:37:25 - meta.AttributeSelectedClassifier
16:37:38 - meta.AttributeSelectedClassifier
16:37:55 - meta.AttributeSelectedClassifier
16:39:55 - meta.AttributeSelectedClassifier

Classifier output

Information Gain Ranking Filter

Ranked attributes:

0.0992	64	x.age80
0.0972	62	x.age.g
0.0956	66	x.age5yr
0.0863	22	diffwalk
0.0826	34	genhlth
0.08	2	employl
0.0629	31	physlth
0.0589	95	x.phys14d
0.0522	87	x.rfhlth
0.0518	97	x.hcvu651
0.0515	67	x.age65yr
0.0432	20	pneuvac4
0.0367	27	rmvteth4
0.0318	102	x.exteth3
0.0297	11	marital
0.0285	71	x.incomg
0.0264	69	x.chldcnt
0.0258	3	income2
0.0246	6	children
0.0242	43	checkup1
0.0241	25	diffalon
0.0209	46	chccpdl
0.0204	77	drocdy3.

Selected attributes: 64,62,66,22,34,2,31,95,87,97,67,20,27,102,11,71,69,3,6,43,25,46,77 : 23

SELECTED ATTRIBUTE BASED ON GAINRATIOATTRIBUTEVAL

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose AttributeSelectedClassifier -E "weka.attributeSelection.GainRatioAttributeEval" -S "weka.attributeSelection.Ranker" -T -1.7976931348623157E308 -N 26" -W weka.classifier

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

16:36:24 - meta.AttributeSelectedClassifier
16:36:37 - meta.AttributeSelectedClassifier
16:36:49 - meta.AttributeSelectedClassifier
16:37:01 - meta.AttributeSelectedClassifier
16:37:14 - meta.AttributeSelectedClassifier
16:37:25 - meta.AttributeSelectedClassifier
16:37:38 - meta.AttributeSelectedClassifier
16:37:55 - meta.AttributeSelectedClassifier

Classifier output

Ranked attributes:

0.1338	22	diffwalk
0.0872	24	diffdres
0.0787	87	x.rfhlth
0.0673	2	employl
0.0651	25	diffalon
0.0579	64	x.age80
0.057	36	chkcdny1
0.053	67	x.age65yr
0.0488	66	x.age5yr
0.0487	46	chccpdl
0.0474	31	physlth
0.0429	97	x.hcvu651
0.0424	95	x.phys14d
0.0413	62	x.age.g
0.0399	34	genhlth
0.0372	53	cvdcrhs4
0.0362	104	x.michd
0.0336	20	pneuvac4
0.033	8	blind
0.0316	50	cvdstrk3
0.0308	13	deaf
0.0302	6	children
0.0288	102	x.exteth3
0.0259	3	income2

Selected attributes: 22,24,87,2,25,64,36,67,66,46,31,97,95,62,34,53,104,20,8,50,13,6,102,3 : 24

SELECTED ATTRIBUTE BASED ON ONERATTRIBUTEVAL

The screenshot shows the Weka interface with the 'Classify' tab selected. A command line at the top reads: `Choose AttributeSelectedClassifier -E "weka.attributeSelection.OneRAttributeEval -S 1 -F 10 -B 6" -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 30" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2`. On the left, there's a configuration panel for 'Test options' with 'Use training set' checked. In the center, under 'Classifier output', a list of selected attributes is displayed:

Attribute Index	Attribute Name
22	diffwalk
34	genhlth
87	x_rfhlth
95	x_phys14d
31	physlth
2	employl
25	diffalon
46	chccopd1
24	diffdres
36	chckdnly1
64	x_age80
53	cvdcrhd4
104	x_michd
62	x_age.g
97	x_hcvu651
67	x.age65yr
11	marital
13	deaf
66	x.ageg5yr
29	diabete3
8	blind
52	cvdinfr4
50	cvdstrk3
48	chcsncr
45	choocncr
98	x_totinda
44	exerany2
51	sleptiml
10	sexl

At the bottom, a box contains the text: `Selected attributes: 22,34,87,95,31,2,25,46,24,36,64,53,104,62,97,67,11,13,66,29,8,52,50,48,45,98,44,51,10 : 29`.

TRAINING

CLASSIFIERS USED

- Naive Bayes
- Logistic
- Support Vector Machine
- J48 Decision Tree
- Bagging (Random Forest)

NAIVE BAYES: Naive Bayes is a supervised machine learning algorithm that is based on the Bayes Theorem with an assumption that all the features that predict the target value are independent of each other. It calculates the probability of each class and then picks the one with the highest probability.

LOGISTIC REGRESSION: Logistic regression is a supervised classification algorithm, and it is used in predictive analytics. Logistic regression estimates the probability of an event occurring, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied to the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\text{Logit}(\pi) = \ln(\pi / (1 - \pi))$$

BAGGING: Bagging, also known as Bootstrap aggregating, is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model. Bagging avoids overfitting of data and is used for both regression and classification models, specifically for decision tree algorithms. Random Forest is an improvement over the bagged decision tree.

SVM: Support Vector Machine or SVM is a Supervised Machine Learning algorithm, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it's common to have access to a dataset of at most a couple of thousands of tagged samples.

DECISION TREE: A decision tree is a supervised machine learning algorithm that is used for classification. This algorithm generates the outcome as the optimized result based upon the tree structure with the conditions or rules. The decision tree algorithm is associated with three major components as Decision Nodes, Design Links, and Decision Leaves. It operates with the Splitting, pruning, and tree selection process. It supports both numerical and categorical data to construct the decision tree. Decision tree algorithms are efficient for large datasets with less time complexity.

TEST RESULTS OF ALL 25 MODELS.

CORRELATION ATTRIBUTE SELECTION WITH NAIVE BAYES ALGORITHM

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier Choose **NaiveBayes**

Test options Use training set Supplied test set Set... Cross-validation Folds 10 Percentage split % 66 More options...

(Nom) havarth3 Start Stop

Result list (right-click for options)

- 15:37:32 - meta.AttributeSelectedClassifier
- 15:37:46 - meta.AttributeSelectedClassifier
- 15:38:00 - meta.AttributeSelectedClassifier
- 15:38:13 - meta.AttributeSelectedClassifier
- 15:38:27 - meta.AttributeSelectedClassifier
- 15:38:40 - meta.AttributeSelectedClassifier
- 15:39:45 - meta.AttributeSelectedClassifier
- 15:39:57 - meta.AttributeSelectedClassifier
- 15:40:10 - meta.AttributeSelectedClassifier
- 15:40:23 - meta.AttributeSelectedClassifier
- 15:49:30 - bayes.NaiveBayes
- 15:50:47 - bayes.NaiveBayes

Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.03 seconds

==== Summary ====
Correctly Classified Instances      920          73.0159 %
Incorrectly Classified Instances   340          26.9841 %
Kappa statistic                      0.4444
Mean absolute error                  0.279
Root mean squared error              0.4889
Relative absolute error              61.1408 %
Root relative squared error         102.0602 %
Total Number of Instances           1260

==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area  Class
          0.718   0.247    0.840     0.718    0.774     0.453   0.791    0.865      2
          0.753   0.282    0.596     0.753    0.665     0.453   0.791    0.651      1
Weighted Avg.                   0.730   0.260    0.753     0.730    0.735     0.453   0.791    0.789

==== Confusion Matrix ====
      a   b   <-- classified as
582 229 |   a = 2
111 338 |   b = 1
```

CORRELATION ATTRIBUTE SELECTION WITH J48 DECISION TREE ALGORITHM.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier Choose **J48 -C 0.25 -M 2**

Test options Use training set Supplied test set Set... Cross-validation Folds 10 Percentage split % 66 More options...

(Nom) havarth3 Start Stop

Result list (right-click for options)

- 15:37:32 - meta.AttributeSelectedClassifier
- 15:37:46 - meta.AttributeSelectedClassifier
- 15:38:00 - meta.AttributeSelectedClassifier
- 15:38:13 - meta.AttributeSelectedClassifier
- 15:38:27 - meta.AttributeSelectedClassifier
- 15:38:40 - meta.AttributeSelectedClassifier
- 15:39:45 - meta.AttributeSelectedClassifier
- 15:39:57 - meta.AttributeSelectedClassifier
- 15:40:10 - meta.AttributeSelectedClassifier
- 15:40:23 - meta.AttributeSelectedClassifier
- 15:49:30 - bayes.NaiveBayes
- 15:50:47 - bayes.NaiveBayes
- 15:53:10 - trees.J48
- 15:54:10 - trees.J48

Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.01 seconds

==== Summary ====
Correctly Classified Instances      906          71.9048 %
Incorrectly Classified Instances   354          28.0952 %
Kappa statistic                      0.3431
Mean absolute error                  0.349
Root mean squared error              0.4473
Relative absolute error              76.4942 %
Root relative squared error         93.3926 %
Total Number of Instances           1260

==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area  Class
          0.866   0.546    0.741     0.866    0.799     0.355   0.737    0.799      2
          0.454   0.134    0.652     0.454    0.535     0.355   0.737    0.573      1
Weighted Avg.                   0.719   0.399    0.709     0.719    0.705     0.355   0.737    0.719

==== Confusion Matrix ====
      a   b   <-- classified as
702 109 |   a = 2
245 204 |   b = 1
```

CORRELATION ATTRIBUTE SELECTION WITH LOGISTIC ALGORITHMS.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier
Choose **Logistic** -R 1.0E-8 -M -1 -num-decimal-places 4

Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66

(Nom) havarth3

Result list (right-click for options)
 15:38:00 - meta.AttributeSelectedClassifier
 15:38:13 - meta.AttributeSelectedClassifier
 15:38:27 - meta.AttributeSelectedClassifier
 15:38:40 - meta.AttributeSelectedClassifier
 15:39:45 - meta.AttributeSelectedClassifier
 15:39:57 - meta.AttributeSelectedClassifier
 15:40:10 - meta.AttributeSelectedClassifier
 15:40:23 - meta.AttributeSelectedClassifier
 15:49:30 - bayes.NaiveBayes
 15:50:47 - bayes.NaiveBayes
 15:53:10 - trees.J48
 15:54:10 - trees.J48
 15:55:00 - trees.RandomForest
 15:55:54 - trees.RandomForest
 15:57:21 - functions.Logistic
 15:58:09 - functions.Logistic

Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.03 seconds
==== Summary ====
Correctly Classified Instances      958          76.0317 %
Incorrectly Classified Instances   302          23.9683 %
Kappa statistic                   0.4354
Mean absolute error               0.3377
Root mean squared error           0.4144
Relative absolute error           73.9995 %
Root relative squared error      86.5208 %
Total Number of Instances        1260

==== Detailed Accuracy By Class ====

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC     ROC Area  PRC Area  Class
0.906   0.503    0.765     0.906    0.830     0.454    0.798    0.852     2
0.497   0.094    0.746     0.497     0.596     0.454    0.798    0.667     1
Weighted Avg.                     0.760     0.357     0.758     0.760     0.746     0.454    0.798    0.786

==== Confusion Matrix ====

      a     b  <-- classified as
735   76 |  a = 2
226  223 |  b = 1
```

CORRELATION ATTRIBUTE SELECTION WITH SUPPORT VECTOR MACHINE ALGORITHM.

weka explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier
Choose **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 -model "C:\\Program Files\\Weka-3-8-6" -seed 1

Test options
 Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66

(Nom) havarth3

Result list (right-click for options)
 21:49:25 - functions.LibSVM
 21:49:47 - functions.LibSVM
 21:52:09 - functions.LibSVM
 21:52:59 - functions.LibSVM

Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.35 seconds
==== Summary ====
Correctly Classified Instances      912          72.381 %
Incorrectly Classified Instances   348          27.619 %
Kappa statistic                   0.3038
Mean absolute error               0.2762
Root mean squared error           0.5255
Relative absolute error           60.5282 %
Root relative squared error      109.7169 %
Total Number of Instances        1260

==== Detailed Accuracy By Class ====

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC     ROC Area  PRC Area  Class
0.953   0.690    0.714     0.953    0.816     0.362    0.631    0.710     2
0.310   0.047    0.785     0.310    0.444     0.362    0.631    0.489     1
Weighted Avg.                     0.724     0.461     0.739     0.724     0.684     0.362    0.631    0.632

==== Confusion Matrix ====

      a     b  <-- classified as
773   38 |  a = 2
310  139 |  b = 1
```

CORRELATION ATTRIBUTE SELECTION WITH BAGGING (RANDOM FOREST) ALGORITHM.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.RandomForest -- -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66
 More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

- 21:49:25 - functions.LibSVM
- 21:49:47 - functions.LibSVM
- 21:52:09 - functions.LibSVM
- 21:52:59 - functions.LibSVM
- 21:54:27 - meta.Bagging
- 21:55:03 - meta.Bagging**

Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 1.28 seconds
==== Summary ====
Correctly Classified Instances      896          71.1111 %
Incorrectly Classified Instances    364          28.8889 %
Kappa statistic                   0.3407
Mean absolute error               0.3452
Root mean squared error           0.4326
Relative absolute error            75.6419 %
Root relative squared error       90.3175 %
Total Number of Instances         1260

==== Detailed Accuracy By Class ====
      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
          0.831    0.506    0.748     0.831    0.787     0.345    0.766    0.849      2
          0.494    0.169    0.618     0.494    0.550     0.345    0.766    0.614      1
Weighted Avg.      0.711    0.386    0.702     0.711    0.703     0.345    0.766    0.765

==== Confusion Matrix ====
      a   b   <-- classified as
674 137 |   a = 2
227 222 |   b = 1
```

CHI-SQUARED ATTRIBUTE SELECTION WITH NAIVE BAYES ALGORITHM:

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **NaiveBayes**

Test options

Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66
 More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

- 09:36:22 - functions.Logistic
- 09:37:05 - functions.Logistic
- 09:37:57 - bayes.NaiveBayes
- 09:38:38 - bayes.NaiveBayes**

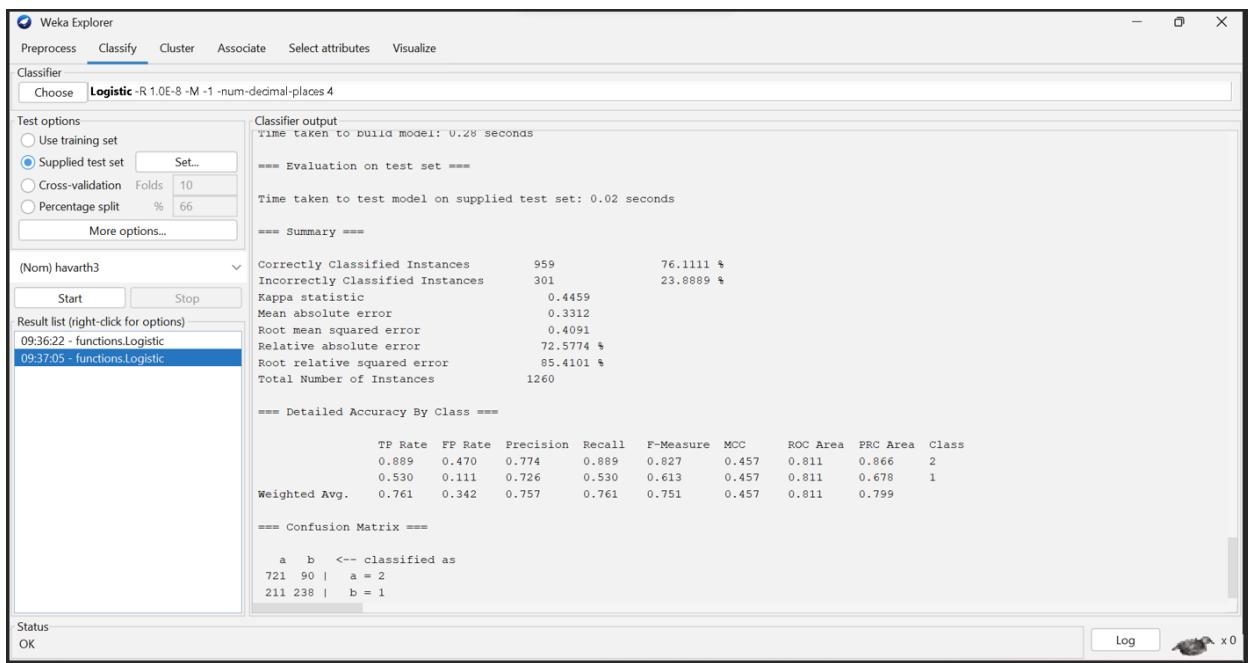
Classifier output

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.01 seconds
==== Summary ====
Correctly Classified Instances      922          73.1746 %
Incorrectly Classified Instances    338          26.8254 %
Kappa statistic                   0.4455
Mean absolute error               0.2765
Root mean squared error           0.4859
Relative absolute error            60.6066 %
Root relative squared error       101.4378 %
Total Number of Instances         1260

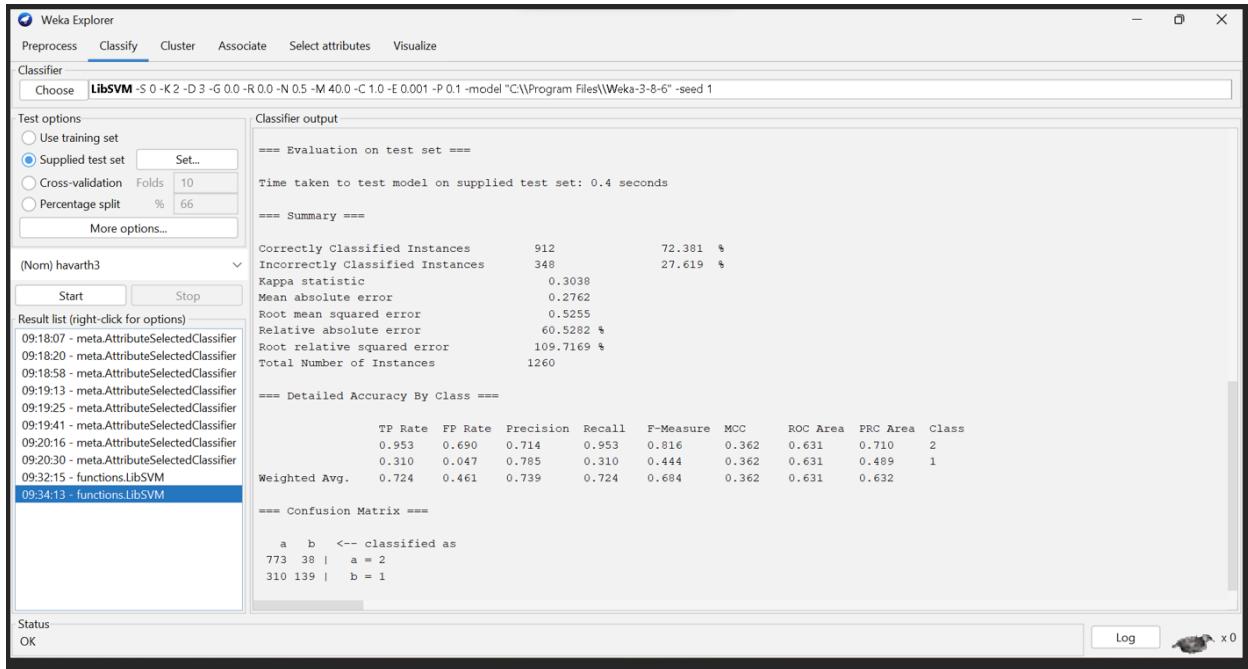
==== Detailed Accuracy By Class ====
      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
          0.724    0.254    0.837     0.724    0.776     0.453    0.796    0.872      2
          0.746    0.276    0.599     0.746    0.665     0.453    0.796    0.648      1
Weighted Avg.      0.732    0.262    0.753     0.732    0.737     0.453    0.796    0.792

==== Confusion Matrix ====
      a   b   <-- classified as
587 224 |   a = 2
114 335 |   b = 1
```

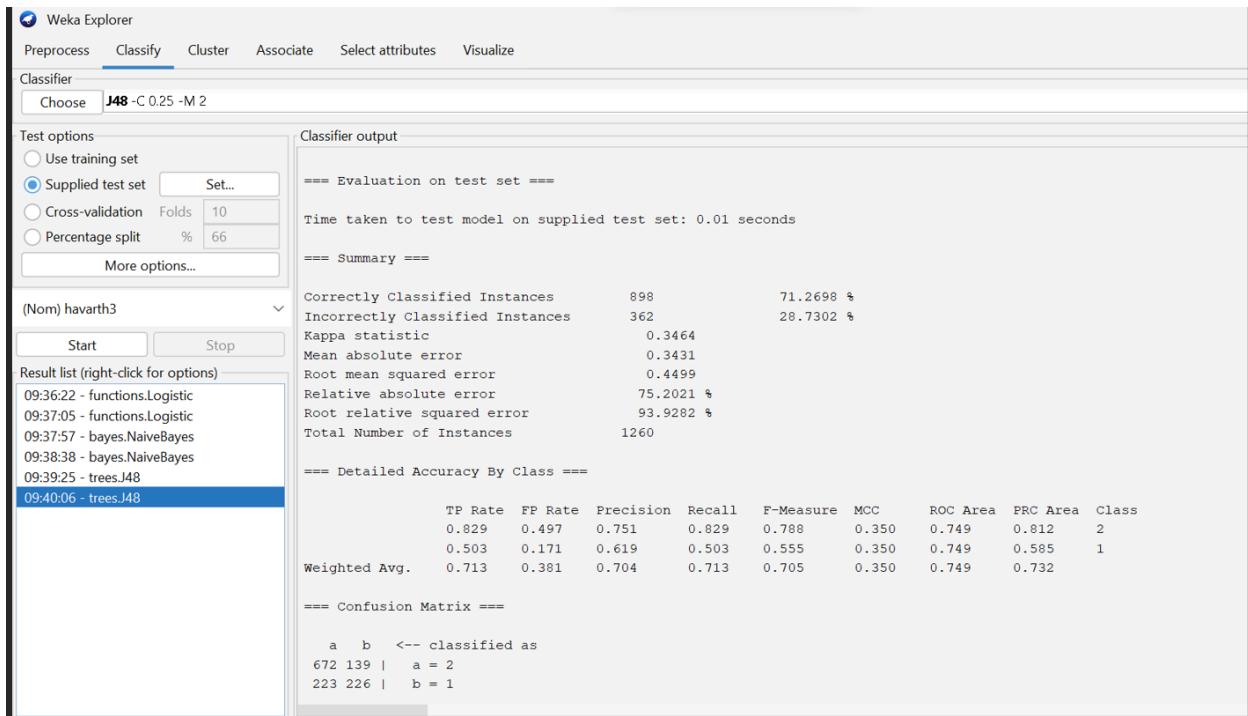
CHI-SQUARED ATTRIBUTE SELECTION WITH LOGISTIC ALGORITHM



CHI-SQUARED ATTRIBUTE SELECTION WITH SUPPORT VECTOR MACHINE ALGORITHM



CHI-SQUARED ATTRIBUTE SELECTION WITH J48 DECISION TREE ALGORITHM



CHI-SQUARED ATTRIBUTE SELECTION WITH BAGGING (RANDOM FOREST) ALGORITHM

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'Bagging' is chosen with the command: `-P 100 -S 1 -num-slots 1 -l 10 -W weka.classifiers.trees.RandomForest -- -P 100 -l 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1`. The 'Test options' section shows 'Supplied test set' selected. The 'Classifier output' pane displays evaluation metrics and a confusion matrix.

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 1.06 seconds
    === Summary ===
    Correctly Classified Instances      901      71.5079 %
    Incorrectly Classified Instances   359      28.4921 %
    Kappa statistic                   0.3508
    Mean absolute error              0.3479
    Root mean squared error          0.4317
    Relative absolute error          76.2328 %
    Root relative squared error     90.1298 %
    Total Number of Instances        1260

    === Detailed Accuracy By Class ===

    TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
    0.832   0.497   0.752     0.832   0.790     0.355   0.769   0.859     2
    0.503   0.168   0.624     0.503   0.557     0.355   0.769   0.608     1
    Weighted Avg.       0.715   0.379   0.706     0.715   0.707     0.355   0.769   0.769

    === Confusion Matrix ===

    a   b   <-- classified as
    675 136 |  a = 2
    223 226 |  b = 1
  
```

INFO GAIN ATTRIBUTE SELECTION WITH NAIVE BAYES ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'NaiveBayes' is chosen. The 'Test options' section shows 'Supplied test set' selected. The 'Classifier output' pane displays evaluation metrics and a confusion matrix.

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0.02 seconds
    === Summary ===
    Correctly Classified Instances      925      73.4127 %
    Incorrectly Classified Instances   335      26.5873 %
    Kappa statistic                   0.4512
    Mean absolute error              0.274
    Root mean squared error          0.4846
    Relative absolute error          60.0428 %
    Root relative squared error     101.1601 %
    Total Number of Instances        1260

    === Detailed Accuracy By Class ===

    TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
    0.724   0.247   0.841     0.724   0.778     0.459   0.798   0.872     2
    0.753   0.276   0.601     0.753   0.669     0.459   0.798   0.653     1
    Weighted Avg.       0.734   0.258   0.756     0.734   0.739     0.459   0.798   0.794

    === Confusion Matrix ===

    a   b   <-- classified as
    587 224 |  a = 2
    111 338 |  b = 1
  
```

INFO GAIN ATTRIBUTE SELECTION WITH LOGISTIC ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'Logistic' is chosen with parameters: -R 1.0E-8 -M 1 -num-decimal-places 4. In the 'Test options' section, 'Supplied test set' is selected with 10 folds. The 'Classifier output' pane displays the evaluation results:

```

time taken to build model: 0.29 seconds
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.01 seconds
==== Summary ====
Correctly Classified Instances      962      76.3492 %
Incorrectly Classified Instances   298      23.6508 %
Kappa statistic                      0.45
Mean absolute error                  0.3301
Root mean squared error              0.4097
Relative absolute error              72.3493 %
Root relative squared error         85.5275 %
Total Number of Instances           1260
==== Detailed Accuracy By Class ====
      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
      0.894     0.472     0.774     0.894     0.830     0.463     0.809     0.864     2
      0.528     0.106     0.734     0.528     0.614     0.463     0.809     0.678     1
Weighted Avg.      0.763     0.342     0.759     0.763     0.753     0.463     0.809     0.797
==== Confusion Matrix ====
      a     b  <-- classified as
    725   86 |  a = 2
    212  237 |  b = 1
  
```

INFO GAIN ATTRIBUTE SELECTION WITH J48 DECISION TREE ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'J48' is chosen with parameters: -C 0.25 -M 2. In the 'Test options' section, 'Supplied test set' is selected with 10 folds. The 'Classifier output' pane displays the evaluation results:

```

==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.01 seconds
==== Summary ====
Correctly Classified Instances      891      70.7143 %
Incorrectly Classified Instances   369      29.2857 %
Kappa statistic                      0.3271
Mean absolute error                  0.3624
Root mean squared error              0.4648
Relative absolute error              79.4161 %
Root relative squared error         97.0434 %
Total Number of Instances           1260
==== Detailed Accuracy By Class ====
      TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
      0.836     0.526     0.742     0.836     0.786     0.333     0.711     0.783     2
      0.474     0.164     0.616     0.474     0.536     0.333     0.711     0.541     1
Weighted Avg.      0.707     0.397     0.697     0.707     0.697     0.333     0.711     0.697
==== Confusion Matrix ====
      a     b  <-- classified as
    678  133 |  a = 2
    236  213 |  b = 1
  
```

INFO GAIN ATTRIBUTE SELECTION WITH SUPPORT VECTOR MACHINE ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'LibSVM' with parameters: -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 -model "C:\Program Files\Weka-3-8-6" -seed 1. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' pane displays evaluation metrics and a confusion matrix.

Classifier output:

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.41 seconds

==== Summary ====
Correctly Classified Instances      912          72.381 %
Incorrectly Classified Instances   348          27.619 %
Kappa statistic                      0.3038
Mean absolute error                  0.2762
Root mean squared error              0.5255
Relative absolute error               60.5282 %
Root relative squared error         109.7169 %
Total Number of Instances            1260

==== Detailed Accuracy By Class ====

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.953	0.690	0.714	0.953	0.816	0.362	0.631	0.710	2	
0.310	0.047	0.785	0.310	0.444	0.362	0.631	0.489	1	
Weighted Avg.	0.724	0.461	0.739	0.724	0.684	0.362	0.631	0.632	

```
==== Confusion Matrix ====
a   b   <-- classified as
773 38 |  a = 2
310 139 |  b = 1
```

INFO GAIN ATTRIBUTE SELECTION WITH BAGGING (RANDOM FOREST) ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'Bagging' with parameters: -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.RandomForest -- -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' pane displays evaluation metrics and a confusion matrix.

Classifier output:

```
==== Evaluation on test set ====
Time taken to test model on supplied test set: 1.05 seconds

==== Summary ====
Correctly Classified Instances      910          72.2222 %
Incorrectly Classified Instances   350          27.7778 %
Kappa statistic                      0.3635
Mean absolute error                  0.3468
Root mean squared error              0.4244
Relative absolute error               76.0095 %
Root relative squared error         88.6109 %
Total Number of Instances            1260

==== Detailed Accuracy By Class ====

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.845	0.499	0.754	0.845	0.797	0.369	0.782	0.870	2	
0.501	0.155	0.641	0.501	0.563	0.369	0.782	0.626	1	
Weighted Avg.	0.722	0.376	0.713	0.722	0.713	0.369	0.782	0.783	

```
==== Confusion Matrix ====
a   b   <-- classified as
685 126 |  a = 2
224 225 |  b = 1
```

GAIN RATIO ATTRIBUTE SELECTION WITH NAIVE BAYES ALGORITHM:

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'NaiveBayes' is chosen. In the 'Test options' section, 'Supplied test set' is selected. The 'Classifier output' pane displays the following evaluation metrics:

```

Time taken to build model: 0.01 seconds
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.03 seconds
==== Summary ====
Correctly Classified Instances      921      73.0952 %
Incorrectly Classified Instances   339      26.9048 %
Kappa statistic                      0.4421
Mean absolute error                  0.2739
Root mean squared error              0.4807
Relative absolute error              60.03 %
Root relative squared error         100.3656 %
Total Number of Instances           1260
==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
      0.727    0.263    0.833    0.727    0.777    0.449  0.801    0.874    2
      0.737    0.273    0.600    0.737    0.661    0.449  0.801    0.666    1
Weighted Avg.          0.731    0.266    0.750    0.731    0.736    0.449  0.801    0.799
==== Confusion Matrix ====
      a   b   <-- classified as
  590 221 |   a = 2
 118 331 |   b = 1
  
```

GAIN RATIO ATTRIBUTE SELECTION WITH LOGISTIC ALGORITHM:

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'Logistic' is chosen with parameters '-R 1.0E-8 -M -1 -num-decimal-places 4'. In the 'Test options' section, 'Supplied test set' is selected. The 'Classifier output' pane displays the following evaluation metrics:

```

Time taken to build model: 0.01 seconds
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.02 seconds
==== Summary ====
Correctly Classified Instances      961      76.2698 %
Incorrectly Classified Instances   299      23.7302 %
Kappa statistic                      0.4431
Mean absolute error                  0.3284
Root mean squared error              0.4062
Relative absolute error              71.9664 %
Root relative squared error         84.8056 %
Total Number of Instances           1260
==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
      0.904    0.492    0.768    0.904    0.831    0.460  0.820    0.881    2
      0.508    0.096    0.745    0.508    0.604    0.460  0.820    0.683    1
Weighted Avg.          0.763    0.351    0.760    0.763    0.750    0.460  0.820    0.810
==== Confusion Matrix ====
      a   b   <-- classified as
  733 78 |   a = 2
 221 228 |   b = 1
  
```

GAIN RATIO ATTRIBUTE SELECTION WITH J48 DECISION TREE ALGORITHM:

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -C 0.25 -M 2'. The 'Test options' section indicates 'Supplied test set' is selected. The 'Classifier output' pane displays the evaluation results on the test set, including summary statistics like Kappa statistic (0.3031), detailed accuracy by class, and a confusion matrix.

```

Classifier output
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.01 seconds

==== Summary ====
Correctly Classified Instances      884          70.1587 %
Incorrectly Classified Instances   376          29.8413 %
Kappa statistic                   0.3031
Mean absolute error               0.3476
Root mean squared error           0.4634
Relative absolute error            76.1874 %
Root relative squared error       96.7352 %
Total Number of Instances         1260

==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
      0.851    0.568    0.730     0.851    0.786     0.313    0.719     0.792     2
      0.432    0.149    0.616     0.432    0.508     0.313    0.719     0.549     1
Weighted Avg.      0.702    0.419    0.689     0.702    0.687     0.313    0.719     0.705

==== Confusion Matrix ====
      a   b   <-- classified as
690 121 |  a = 2
255 194 |  b = 1
  
```

GAIN RATIO ATTRIBUTE SELECTION WITH SUPPORT VECTOR MACHINE ALGORITHM:

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is '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 -model "C:\Program Files\Weka-3-8-6" -seed 1'. The 'Test options' section indicates 'Supplied test set' is selected. The 'Classifier output' pane displays the evaluation results on the test set, including summary statistics like Kappa statistic (0.3038), detailed accuracy by class, and a confusion matrix.

```

Classifier output
==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.35 seconds

==== Summary ====
Correctly Classified Instances      912          72.381 %
Incorrectly Classified Instances   348          27.619 %
Kappa statistic                   0.3038
Mean absolute error               0.2762
Root mean squared error           0.5255
Relative absolute error            60.5282 %
Root relative squared error       109.7169 %
Total Number of Instances         1260

==== Detailed Accuracy By Class ====
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
      0.953    0.690    0.714     0.953    0.816     0.362    0.631    0.710     2
      0.310    0.047    0.785     0.310    0.444     0.362    0.631    0.489     1
Weighted Avg.      0.724    0.461    0.739     0.724    0.684     0.362    0.631    0.632

==== Confusion Matrix ====
      a   b   <-- classified as
773  38 |  a = 2
310 139 |  b = 1
  
```

GAIN RATIO ATTRIBUTE SELECTION WITH BAGGING (RANDOM FOREST) ALGORITHM:

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'Bagging' is chosen with the command: `-P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.RandomForest -- -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1`. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' pane displays evaluation metrics and a confusion matrix.

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 1.39 seconds
    === Summary ===
    Correctly Classified Instances      909      72.1429 %
    Incorrectly Classified Instances   351      27.8571 %
    Kappa statistic                   0.375
    Mean absolute error              0.3391
    Root mean squared error          0.4279
    Relative absolute error          74.3242 %
    Root relative squared error     89.3291 %
    Total Number of Instances        1260

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.819     0.454     0.765      0.819     0.791      0.377   0.777     0.860      2
    0.546     0.181     0.625      0.546     0.583      0.377   0.777     0.618      1
    Weighted Avg.   0.721     0.357     0.715      0.721     0.717      0.377   0.777     0.773

    === Confusion Matrix ===
    a   b   <-- classified as
    664 147 |   a = 2
    204 245 |   b = 1
  
```

ONE R ATTRIBUTE SELECTION WITH NAIVE BAYES ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. Under 'Classifier', 'NaiveBayes' is chosen. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' pane displays evaluation metrics and a confusion matrix.

```

    === Evaluation on test set ===
    Time taken to test model on supplied test set: 0.02 seconds
    === Summary ===
    Correctly Classified Instances      925      73.4127 %
    Incorrectly Classified Instances   335      26.5873 %
    Kappa statistic                   0.4396
    Mean absolute error              0.2707
    Root mean squared error          0.4702
    Relative absolute error          59.3247 %
    Root relative squared error     98.1727 %
    Total Number of Instances        1260

    === Detailed Accuracy By Class ===
    TP Rate   FP Rate   Precision   Recall   F-Measure   MCC   ROC Area   PRC Area   Class
    0.751     0.296     0.821      0.751     0.784      0.443   0.803     0.875      2
    0.704     0.249     0.610      0.704     0.654      0.443   0.803     0.668      1
    Weighted Avg.   0.734     0.279     0.746      0.734     0.738      0.443   0.803     0.801

    === Confusion Matrix ===
    a   b   <-- classified as
    609 202 |   a = 2
    133 316 |   b = 1
  
```

ONE R ATTRIBUTE SELECTION WITH LOGISTIC ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'Logistic -R 1.0E-8 -M -1 -num-decimal-places 4'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the evaluation results for the test set:

```

==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.01 seconds

==== Summary ====
Correctly Classified Instances      951      75.4762 %
Incorrectly Classified Instances   309      24.5238 %
Kappa statistic                      0.43
Mean absolute error                  0.3254
Root mean squared error              0.4076
Relative absolute error              71.323 %
Root relative squared error         85.0861 %
Total Number of Instances           1260

```

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.887	0.483	0.768	0.887	0.823	0.442	0.820	0.881	0.881	2
0.517	0.113	0.716	0.517	0.600	0.442	0.820	0.677	0.677	1
Weighted Avg.	0.755	0.351	0.750	0.755	0.744	0.442	0.820	0.808	

==== Confusion Matrix ====

	a	b	-- classified as
a	719	92	a = 2
b	217	232	b = 1

ONE R ATTRIBUTE SELECTION WITH J48 DECISION TREE ALGORITHM.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -C 0.25 -M 2'. The 'Test options' panel shows 'Supplied test set' selected. The 'Classifier output' panel displays the evaluation results for the test set:

```

==== Evaluation on test set ====
Time taken to build model: 0.14 seconds
Time taken to test model on supplied test set: 0.01 seconds

==== Summary ====
Correctly Classified Instances      914      72.5397 %
Incorrectly Classified Instances   346      27.4603 %
Kappa statistic                      0.3948
Mean absolute error                  0.342
Root mean squared error              0.4574
Relative absolute error              74.954 %
Root relative squared error         95.4888 %
Total Number of Instances           1260

```

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.800	0.410	0.779	0.800	0.790	0.395	0.718	0.782	0.782	2
0.590	0.200	0.621	0.590	0.605	0.395	0.718	0.568	0.568	1
Weighted Avg.	0.725	0.335	0.723	0.725	0.724	0.395	0.718	0.706	

==== Confusion Matrix ====

	a	b	-- classified as
a	649	162	a = 2
b	184	265	b = 1

ONE R ATTRIBUTE SELECTION WITH SUPPORT VECTOR MACHINE ALGORITHM.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **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 -model "C:\Program Files\Weka-3-8-6" -seed 1

Test options

Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66
 More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

21:49:25 - functions.LibSVM
21:49:47 - functions.LibSVM

Classifier output

==== Evaluation on test set ====
Time taken to test model on supplied test set: 0.45 seconds
==== Summary ====

	Correctly Classified Instances	914	72.5397 %
Incorrectly Classified Instances	346	27.4603 %	
Kappa statistic	0.3094		
Mean absolute error	0.2746		
Root mean squared error	0.524		
Relative absolute error	60.1803 %		
Root relative squared error	109.4012 %		
Total Number of Instances	1260		

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.952	0.684	0.715	0.952	0.817	0.366	0.634	0.712	2	
0.316	0.048	0.785	0.316	0.451	0.366	0.634	0.492	1	
Weighted Avg.	0.725	0.457	0.740	0.725	0.686	0.366	0.634	0.634	

==== Confusion Matrix ====

	a	b	<-- classified as
772	39	1	a = 2
307	142	1	b = 1

ONE R ATTRIBUTE SELECTION WITH BAGGING (RANDOM FOREST) ALGORITHM.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **Bagging** -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.RandomForest -- -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

Use training set
 Supplied test set Set...
 Cross-validation Folds 10
 Percentage split % 66
 More options...

(Nom) havarth3

Start Stop

Result list (right-click for options)

17:32:28 - bayes.NaiveBayes
17:33:34 - bayes.NaiveBayes
21:02:57 - functions.Logistic
21:05:01 - functions.Logistic
21:05:59 - trees.J48
21:06:44 - trees.J48
21:07:22 - trees.RandomForest
21:08:03 - meta.Bagging
21:09:05 - meta.Bagging

Classifier output

==== Evaluation on test set ====
Time taken to test model on supplied test set: 1.28 seconds
==== Summary ====

	Correctly Classified Instances	897	71.1905 %
Incorrectly Classified Instances	363	28.8095 %	
Kappa statistic	0.3583		
Mean absolute error	0.3476		
Root mean squared error	0.4294		
Relative absolute error	76.1745 %		
Root relative squared error	89.6438 %		
Total Number of Instances	1260		

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.803	0.452	0.762	0.803	0.782	0.359	0.770	0.858	2	
0.548	0.197	0.606	0.548	0.575	0.359	0.770	0.634	1	
Weighted Avg.	0.712	0.361	0.707	0.712	0.708	0.359	0.770	0.778	

==== Confusion Matrix ====

	a	b	<-- classified as
651	160	1	a = 2
203	246	1	b = 1

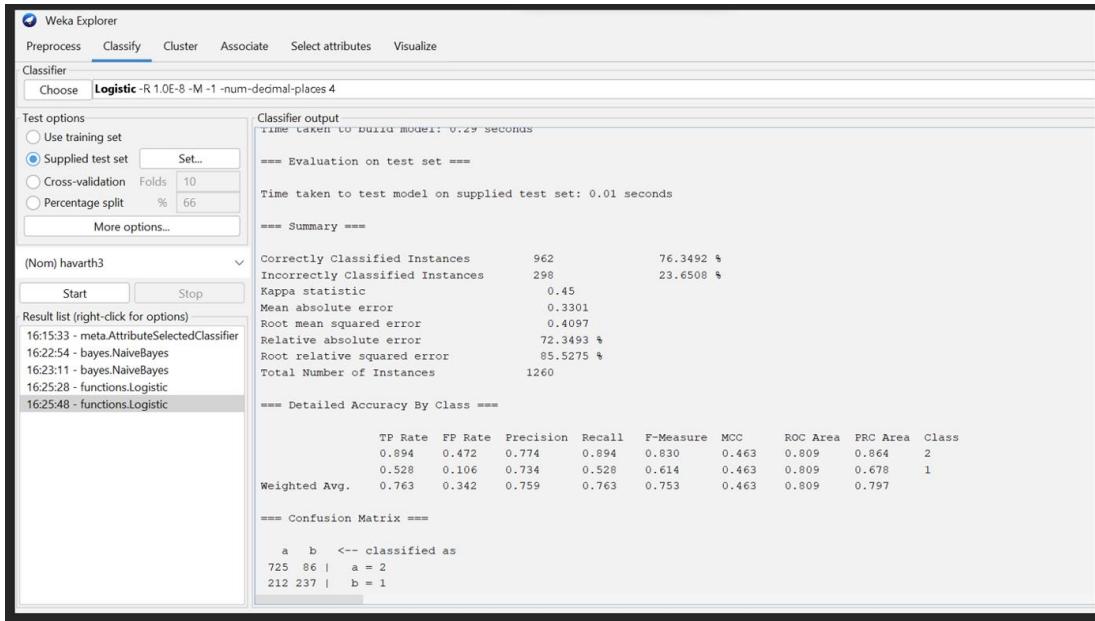
Attribute selection algorithm	Classification algorithm	Accuracy	ROC
Correlation attribute selection	Naive Bayes	73.0159	0.791
	Logistic	76.0317	0.798
	J48 Decision Tree	71.9048	0.737
	SVM	72.381	0.631
	Bagging (Random Forest)	71.1111	0.766
	Classification algorithm	Accuracy	ROC
Chi Squared Attribute selection	Naive Bayes	73.1746	0.796
	Logistic	76.111	0.811
	J48 Decision Tree	72.381	0.631
	SVM	71.2698	0.749
	Bagging (Random Forest)	71.5079	0.769
	Classification algorithm	Accuracy	ROC
Info Gain attribute selection	Naive Bayes	73.4127	0.798
	Logistic	76.3492	0.809
	J48 Decision Tree	70.7143	0.711
	SVM	72.381	0.631
	Bagging (Random Forest)	72.222	0.782
	Classification algorithm	Accuracy	ROC
Gain Ratio Attribute selection	Naive Bayes	73.0952	0.801
	Logistic	76.2698	0.820
	J48 Decision Tree	70.1587	0.719
	SVM	72.381	0.631
	Bagging (Random Forest)	72.1429	0.777
	Classification algorithm	Accuracy	ROC
One R attribute selection	Naive Bayes	73.4127	0.803
	Logistic	75.4762	0.820
	J48 Decision Tree	72.5397	0.718
	SVM	72.5397	0.634
	Bagging (Random Forest)	71.1905	0.770

CONCLUSION

Info Gain attribute selection method with Logistic regression algorithm gives me the best results. Logistic regression gave me the best results for all attribute selection methods, but it achieved the highest accuracy when used with Info Gain. For getting this model, I used ranker = 23. This model used the following attributes:

```
64 x.age80
62 x.age.g
66 x.ageg5yr
22 diffwalk
34 genhlth
2 employl
31 physhlth
95 x.phys14d
87 x.rfhlth
97 x.hcvu651
67 x.age65yr
20 pneuvac4
27 rmvteth4
102 x.exteth3
11 marital
71 x.incomg
69 x.chldcnt
3 income2
6 children
43 checkup1
25 diffalon
46 chccopd1
77 drocdy3.
```

And achieved the following performance on the test data:



FIVE ATTRIBUTES THAT ARE MOST RELEVANT TO THE CLASS ATTRIBUTES:

- x.age80 (Age > 80)
- Diffwalk (Difficulty in walking)
- Genhlth (General Health)
- Chckdny (kidney diseases)
- chccopd1 (chronic obstructive pulmonary disease, emphysema, or chronic bronchitis)

I selected these attributes because all five attributes selection methods suggest them. In general, we also know that age, heart disease, kidney disease, health in general, and difficulty in walking are connected to arthritis.

WHAT I LEARNED FROM THIS PROJECT:

From this project, I got hands-on experience in data mining. I learned how to:

- Handle missing values
- Remove duplicate values
- Standardize data
- Handle outliers
- Split the dataset for training and testing
- Select features
- Implement different classification algorithms
- Compare models

