Assignment 3 EECS 6412

Daniel Marchena Parreira - 216181497

2018/11/05

Question 1

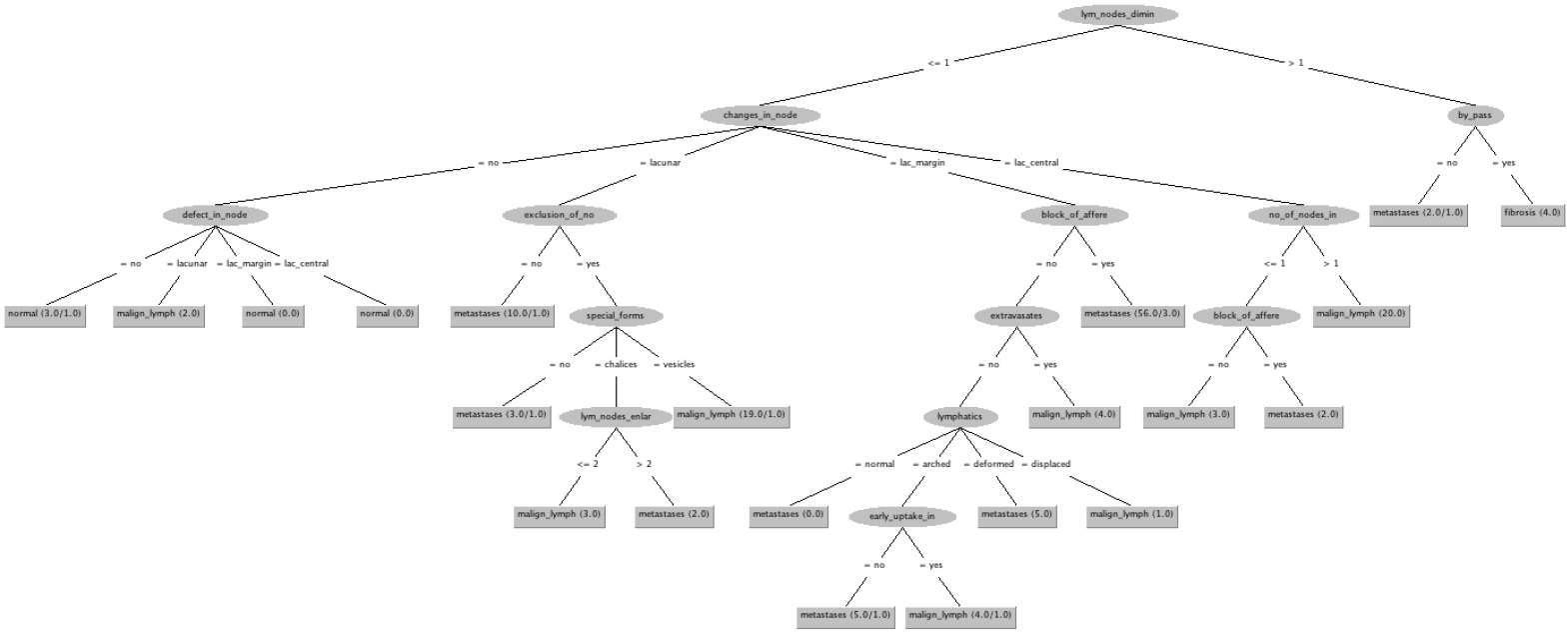
C4.5 (weka.classifier.trees.J48)

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

lym_nodes_dimin <= 1
|   changes_in_node = no
|   |   defect_in_node = no: normal (3.0/1.0)
|   |   defect_in_node = lacunar: malign_lymph (2.0)
|   |   defect_in_node = lac_margin: normal (0.0)
|   |   defect_in_node = lac_central: normal (0.0)
|   changes_in_node = lacunar
|   |   exclusion_of_no = no: metastases (10.0/1.0)
|   |   exclusion_of_no = yes
|   |   |   special_forms = no: metastases (3.0/1.0)
|   |   |   special_forms = chalices
|   |   |   |   lym_nodes_enlar <= 2: malign_lymph (3.0)
|   |   |   |   lym_nodes_enlar > 2: metastases (2.0)
|   |   |   special_forms = vesicles: malign_lymph (19.0/1.0)
|   changes_in_node = lac_margin
|   |   block_of_affere = no
|   |   |   extravasates = no
|   |   |   |   lymphatics = normal: metastases (0.0)
|   |   |   |   lymphatics = arched
|   |   |   |   |   early_uptake_in = no: metastases (5.0/1.0)
|   |   |   |   |   early_uptake_in = yes: malign_lymph (4.0/1.0)
|   |   |   |   lymphatics = deformed: metastases (5.0)
|   |   |   |   lymphatics = displaced: malign_lymph (1.0)
|   |   |   extravasates = yes: malign_lymph (4.0)
|   |   block_of_affere = yes: metastases (56.0/3.0)
|   changes_in_node = lac_central
|   |   no_of_nodes_in <= 1
|   |   |   block_of_affere = no: malign_lymph (3.0)
|   |   |   block_of_affere = yes: metastases (2.0)
|   |   no_of_nodes_in > 1: malign_lymph (20.0)
lym_nodes_dimin > 1
|   by_pass = no: metastases (2.0/1.0)
|   by_pass = yes: fibrosis (4.0)

Number of Leaves  :  21

Size of the tree :  34
```



RIPPER (weka.classifier.rules.JRip)

```
JRIP rules:
===========

(lymphatics = normal) => class=normal (2.0/0.0)
(lym_nodes_dimin >= 2) and (by_pass = yes) => class=fibrosis (4.0/0.0)
(no_of_nodes_in >= 3) and (special_forms = vesicles) => class=malign_lymph (41.0/5.0)
(block_of_affere = no) and (extravasates = yes) => class=malign_lymph (8.0/0.0)
(changes_in_node = lac_central) => class=malign_lymph (8.0/2.0)
 => class=metastases (85.0/11.0)

 Number of Rules : 6
```

Question 2

C4.5 (weka.classifier.trees.J48)

```
=== Summary ===

Correctly Classified Instances         779               96.1728 %
Incorrectly Classified Instances        31                3.8272 %
Kappa statistic                          0.9553
Mean absolute error                      0.0127
Root mean squared error                  0.1005
Relative absolute error                  5.1771 %
Root relative squared error             28.6807 %
Total Number of Instances              810

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 124   0   0   0   1   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   1   0 119   0   2   0   0 |   c = foliage
   1   0   0 107   2   0   0 |   d = cement
   1   0  12   7 105   0   1 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   0   1   0   0   2 120 |   g = grass
```

RIPPER (weka.classifier.rules.JRip)

```
=== Summary ===

Correctly Classified Instances         764               94.321  %
Incorrectly Classified Instances        46                5.679  %
Kappa statistic                          0.9337
Mean absolute error                      0.0243
Root mean squared error                  0.1243
Relative absolute error                  9.9168 %
Root relative squared error             35.4894 %
Total Number of Instances              810

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 122   0   0   0   3   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   0   0 112   2   8   0   0 |   c = foliage
   1   0   0 106   2   1   0 |   d = cement
   1   0  17   7 101   0   0 |   e = window
   0   0   0   1   0  93   0 |   f = path
   1   0   2   0   0   0 120 |   g = grass
```

k-Nearest Neighbor (weka.classifiers.lazy.IBk)

```
=== Summary ===

Correctly Classified Instances         776               95.8025 %
Incorrectly Classified Instances        34                4.1975 %
Kappa statistic                          0.951
Mean absolute error                      0.013
Root mean squared error                  0.1093
Relative absolute error                  5.318  %
Root relative squared error             31.1994 %
Total Number of Instances              810

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 124   0   0   0   1   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   0   0 116   0   6   0   0 |   c = foliage
   1   0   1 103   5   0   0 |   d = cement
   2   0   7   8 109   0   0 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   0   0   1   1   1 120 |   g = grass
```

Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)

```
=== Summary ===

Correctly Classified Instances         624               77.037  %
Incorrectly Classified Instances       186               22.963  %
Kappa statistic                          0.7316
Mean absolute error                      0.0659
Root mean squared error                  0.2464
Relative absolute error                 26.8912 %
Root relative squared error             70.3431 %
Total Number of Instances              810

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 119   0   0   1   5   0   0 |   a = brickface
   0 109   0   1   0   0   0 |   b = sky
   3   0  17   1 101   0   0 |   c = foliage
   9   0   1  97   0   3   0 |   d = cement
  30   0   6  15  75   0   0 |   e = window
   0   0   1   5   0  88   0 |   f = path
   0   0   1   0   3   0 119 |   g = grass
```

Bayesian Network (weka.classifiers.bayes.BayesNet)

```
=== Summary ===

Correctly Classified Instances         753               92.963  %
Incorrectly Classified Instances        57                7.037  %
Kappa statistic                          0.9178
Mean absolute error                      0.0234
Root mean squared error                  0.1288
Relative absolute error                  9.5474 %
Root relative squared error             36.7584 %
Total Number of Instances              810

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 122   0   1   0   2   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   3   0 109   4   6   0   0 |   c = foliage
   4   0   0 100   2   4   0 |   d = cement
   2   0  11  14  99   0   0 |   e = window
   0   0   0   2   0  92   0 |   f = path
   1   0   1   0   0   0 121 |   g = grass
```

Neural Network (weka.classifiers.functions.MultilayerPerceptron)

```
=== Summary ===

Correctly Classified Instances         769               94.9383 %
Incorrectly Classified Instances        41                5.0617 %
Kappa statistic                          0.9409
Mean absolute error                      0.0177
Root mean squared error                  0.1082
Relative absolute error                  7.2247 %
Root relative squared error             30.8905 %
Total Number of Instances              810

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 123   0   0   2   0   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   1   0 109   1  11   0   0 |   c = foliage
   0   0   1 107   1   1   0 |   d = cement
   3   0   8   9 106   0   0 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   0   1   0   0   2 120 |   g = grass
```

Support Vector Machine (weka.classifiers.functions.SMO)

```
=== Summary ===

Correctly Classified Instances         751               92.716  %
Incorrectly Classified Instances        59                7.284  %
Kappa statistic                          0.9149
Mean absolute error                      0.2055
Root mean squared error                  0.3032
Relative absolute error                 83.8368 %
Root relative squared error             86.5505 %
Total Number of Instances              810

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 122   0   1   0   2   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   0   0 111   2   9   0   0 |   c = foliage
   0   0   0 104   5   1   0 |   d = cement
   1   0  20  15  90   0   0 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   0   1   0   0   2 120 |   g = grass
```

Question 3

C4.5 (weka.classifier.trees.J48)

```
Time taken to build model: 0.04 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         230               83.3333 %
Incorrectly Classified Instances        46               16.6667 %
Kappa statistic                          0.665
Mean absolute error                      0.2141
Root mean squared error                  0.3822
Relative absolute error                 43.1117 %
Root relative squared error             76.0455 %
Total Number of Instances              276
```

PRISM (weka.classifiers.rules.Prism)

```
Time taken to build model: 0.09 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         198               71.7391 %
Incorrectly Classified Instances        58               21.0145 %
Kappa statistic                          0.5496
Mean absolute error                      0.2266
Root mean squared error                  0.476
Relative absolute error                 49.3868 %
Root relative squared error             98.7455 %
UnClassified Instances                  20                7.2464 %
Total Number of Instances              276
```

Naive Bayes Classifier (weka.classifiers.bayes.NaiveBayes)

```
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         206               74.6377 %
Incorrectly Classified Instances        70               25.3623 %
Kappa statistic                          0.4843
Mean absolute error                      0.2571
Root mean squared error                  0.4771
Relative absolute error                 51.7659 %
Root relative squared error             94.9347 %
Total Number of Instances              276
```

Bayesian Network (weka.classifiers.bayes.BayesNet)

```
Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         232               84.058  %
Incorrectly Classified Instances        44               15.942  %
Kappa statistic                          0.6783
Mean absolute error                      0.1783
```

```
Root mean squared error                  0.3626
Relative absolute error                 35.91   %
Root relative squared error             72.1546 %
Total Number of Instances              276
```

Neural Network (weka.classifiers.functions.MultilayerPerceptron)

```
Time taken to build model: 4.33 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         231               83.6957 %
Incorrectly Classified Instances        45               16.3043 %
Kappa statistic                          0.6724
Mean absolute error                      0.1751
Root mean squared error                  0.3932
Relative absolute error                 35.2528 %
Root relative squared error             78.2433 %
Total Number of Instances              276
```

Support Vector Machine (weka.classifiers.functions.SMO)

```
Time taken to build model: 0.35 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         224               81.1594 %
Incorrectly Classified Instances        52               18.8406 %
Kappa statistic                          0.6213
Mean absolute error                      0.1884
Root mean squared error                  0.4341
Relative absolute error                 37.9349 %
Root relative squared error             86.3682 %
Total Number of Instances              276
```

Question 4

```
=== Chi-Squared statistic - Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 35 class):
Chi-squared Ranking Filter

Ranked attributes:
194.7609    6 a06
186.5519    5 a05
170.242    34 a34
169.6782   33 a33
169.2576   29 a29

=== Information gain - Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 35 class):
Information Gain Ranking Filter

Ranked attributes:
 0.462    5 a05
 0.44     6 a06
 0.401   33 a33
 0.391   29 a29
 0.384    3 a03
```

C4.5 (weka.classifiers.trees.J48)

```
=== Summary ===

Correctly Classified Instances         321               91.453  %
Incorrectly Classified Instances        30                8.547  %
Kappa statistic                          0.8096
Mean absolute error                      0.0938
Root mean squared error                  0.2901
Relative absolute error                 20.36   %
Root relative squared error             60.4599 %
Total Number of Instances              351

=== Chi-Squared statistic - Summary ===

Correctly Classified Instances         319               90.8832 %
Incorrectly Classified Instances        32                9.1168 %
Kappa statistic                          0.7955
Mean absolute error                      0.1344
Root mean squared error                  0.2911
Relative absolute error                 29.1856 %
Root relative squared error             60.6881 %
Total Number of Instances              351
```

```
=== Information gain - Summary ===

Correctly Classified Instances         329               93.7322 %
Incorrectly Classified Instances        22                6.2678 %
Kappa statistic                          0.8604
Mean absolute error                      0.1084
Root mean squared error                  0.2441
Relative absolute error                 23.5377 %
Root relative squared error             50.8782 %
Total Number of Instances               351
```

RIPPER (weka.classifier.rules.JRip)

```
=== Summary ===

Correctly Classified Instances         315               89.7436 %
Incorrectly Classified Instances        36               10.2564 %
Kappa statistic                          0.7779
Mean absolute error                      0.1447
Root mean squared error                  0.3041
Relative absolute error                 31.4132 %
Root relative squared error             63.3836 %
Total Number of Instances               351

=== Chi-Squared statistic - Summary ===

Correctly Classified Instances         317               90.3134 %
Incorrectly Classified Instances        34                9.6866 %
Kappa statistic                          0.7842
Mean absolute error                      0.1421
Root mean squared error                  0.3006
Relative absolute error                 30.8499 %
Root relative squared error             62.6528 %
Total Number of Instances               351

=== Information gain - Summary ===

Correctly Classified Instances         319               90.8832 %
Incorrectly Classified Instances        32                9.1168 %
Kappa statistic                          0.7991
Mean absolute error                      0.1327
Root mean squared error                  0.292
Relative absolute error                 28.8145 %
Root relative squared error             60.8643 %
Total Number of Instances               351
```

Naive Bayesian Classification (weka.classifiers.bayes.NaiveBayes)

```
=== Summary ===

Correctly Classified Instances         290               82.6211 %
Incorrectly Classified Instances        61               17.3789 %
Kappa statistic                          0.6394
Mean absolute error                      0.1736
Root mean squared error                  0.3935
Relative absolute error                 37.7001 %
Root relative squared error             82.0203 %
Total Number of Instances              351

=== Chi-Squared statistic - Summary ===

Correctly Classified Instances         297               84.6154 %
Incorrectly Classified Instances        54               15.3846 %
Kappa statistic                          0.6669
Mean absolute error                      0.1763
Root mean squared error                  0.3312
Relative absolute error                 38.2865 %
Root relative squared error             69.0352 %
Total Number of Instances              351

=== Information gain - Summary ===

Correctly Classified Instances         310               88.3191 %
Incorrectly Classified Instances        41               11.6809 %
Kappa statistic                          0.743
Mean absolute error                      0.1448
Root mean squared error                  0.3231
Relative absolute error                 31.4437 %
Root relative squared error             67.3524 %
Total Number of Instances              351
```

Bayesian Network (weka.classifiers.bayes.BayesNet)

```
=== Summary ===

Correctly Classified Instances         314               89.4587 %
Incorrectly Classified Instances        37               10.5413 %
Kappa statistic                          0.7681
Mean absolute error                      0.1069
Root mean squared error                  0.3179
Relative absolute error                 23.2213 %
Root relative squared error             66.2607 %
Total Number of Instances              351
```

```
=== Chi-Squared statistic - Summary ===

Correctly Classified Instances         317               90.3134 %
Incorrectly Classified Instances        34                9.6866 %
Kappa statistic                          0.785
Mean absolute error                      0.1087
Root mean squared error                  0.2927
Relative absolute error                 23.5955 %
Root relative squared error             61.0198 %
Total Number of Instances              351

=== Information gain - Summary ===

Correctly Classified Instances         320               91.1681 %
Incorrectly Classified Instances        31                8.8319 %
Kappa statistic                          0.8057
Mean absolute error                      0.0962
Root mean squared error                  0.274
Relative absolute error                 20.8864 %
Root relative squared error             57.1108 %
Total Number of Instances              351
```

k-Nearest Neighbor (weka.classifiers.lazy.IBk)

```
=== Summary ===

Correctly Classified Instances         303               86.3248 %
Incorrectly Classified Instances        48               13.6752 %
Kappa statistic                          0.6841
Mean absolute error                      0.139
Root mean squared error                  0.3686
Relative absolute error                 30.1815 %
Root relative squared error             76.8426 %
Total Number of Instances              351

=== Chi-Squared statistic - Summary ===

Correctly Classified Instances         315               89.7436 %
Incorrectly Classified Instances        36               10.2564 %
Kappa statistic                          0.7724
Mean absolute error                      0.1047
Root mean squared error                  0.3193
Relative absolute error                 22.7451 %
Root relative squared error             66.552  %
Total Number of Instances              351

=== Information gain - Summary ===

Correctly Classified Instances         314               89.4587 %
Incorrectly Classified Instances        37               10.5413 %
Kappa statistic                          0.7681
Mean absolute error                      0.1137
Root mean squared error                  0.3265
Relative absolute error                 24.6828 %
Root relative squared error             68.0565 %
Total Number of Instances              351
```

Neural networks (weka.classifiers.functions.MultilayerPerceptron)

```
=== Summary ===

Correctly Classified Instances         320               91.1681 %
Incorrectly Classified Instances        31                8.8319 %
Kappa statistic                          0.7993
Mean absolute error                      0.0938
Root mean squared error                  0.2786
Relative absolute error                 20.3738 %
Root relative squared error             58.0756 %
Total Number of Instances              351

=== Chi-Squared statistic - Summary ===

Correctly Classified Instances         320               91.1681 %
Incorrectly Classified Instances        31                8.8319 %
Kappa statistic                          0.8022
Mean absolute error                      0.1289
Root mean squared error                  0.283
Relative absolute error                 27.9888 %
Root relative squared error             58.9917 %
Total Number of Instances              351

=== Information gain - Summary ===

Correctly Classified Instances         321               91.453  %
Incorrectly Classified Instances        30                8.547  %
Kappa statistic                          0.811
Mean absolute error                      0.1163
Root mean squared error                  0.274
Relative absolute error                 25.255  %
Root relative squared error             57.1107 %
Total Number of Instances              351
```

Brief discussion:

After running the algorithms using the top five attributes obtained by the attribute selection methods, we notice a slightly positive response in terms of correctly classified instances compared to using all attributes available.

The reason for that is because attribute selection methods are able to rank the most relevant attributes based on how "pure" they can partition a dataset. On that note, that will lead to simpler, faster and more accurate models.

Correctly classified improvement:

- **C4.5**: got 2% more efficient using Information Gain
- **RIPPER**: got 1% more efficient using Information Gain
- **Naive Bayesian Classification**: got 6% more efficient using Information Gain
- **Bayesian Network**: got 2% more efficient using Information Gain
- **k-Nearest Neighbor**: got 3% more efficient using Information Gain
- **Neural networks**: got 0.4% more efficient using Information Gain, but it had a tremendous gain in run time

Question 5

Balance-scale data set

- **C4.5**: 35.52% error rate
- **RIPPER**: 29.92% error rate
- **Naive Bayesian Classification**: 8.64% error rate
- **Bayesian Network**: 8.64% error rate
- **k-Nearest Neighbor**: 16.16% error rate
- **Neural networks**: 2.08% error rate

Ecoli data set

- **C4.5**: 15.77% error rate
- **RIPPER**: 19.34% error rate
- **Naive Bayesian Classification**: 14.58% error rate
- **Bayesian Network**: 18.75% error rate
- **k-Nearest Neighbor**: 19.64% error rate
- **Neural networks**: 13.98% error rate

Glass Identification data set

- **C4.5**: 34.11% error rate
- **RIPPER**: 30.37% error rate
- **Naive Bayesian Classification**: 50.46% error rate
- **Bayesian Network**: 25.23% error rate
- **k-Nearest Neighbor**: 29.43% error rate
- **Neural networks**: 30.84% error rate

Ionosphere data set

- **C4.5**: 8.54% error rate
- **RIPPER**: 10.25% error rate
- **Naive Bayesian Classification**: 17.37% error rate
- **Bayesian Network**: 10.54% error rate
- **k-Nearest Neighbor**: 13.67% error rate
- **Neural networks**: 8.83% error rate

Iris plant data set

- **C4.5**: 4% error rate
- **RIPPER**: 4.66% error rate
- **Naive Bayesian Classification**: 4% error rate
- **Bayesian Network**: 7.33% error rate
- **k-Nearest Neighbor**: 4.66% error rate
- **Neural networks**: 2.66% error rate

Wine data set

- **C4.5**: 6.17% error rate
- **RIPPER**: 7.86% error rate
- **Naive Bayesian Classification**: 3.37% error rate
- **Bayesian Network**: 1.12% error rate
- **k-Nearest Neighbor**: 5.05% error rate
- **Neural networks**: 2.80% error rate

Yeast data set

- **C4.5**: 44.07% error rate
- **RIPPER**: 41.91% error rate
- **Naive Bayesian Classification**: 42.38% error rate
- **Bayesian Network**: 43.26% error rate
- **k-Nearest Neighbor**: 47.70% error rate
- **Neural networks**: 40.56% error rate

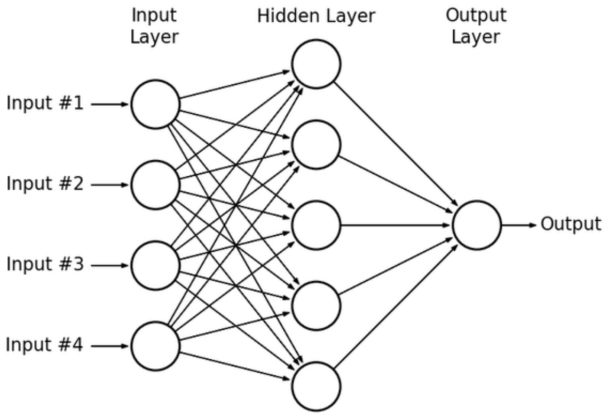
Average classification error rate of each method over all the data sets

- **C4.5**: 21.16% error rate
- **RIPPER**: 20.61% error rate
- **Naive Bayesian Classification**: 20.11% error rate
- **k-Nearest Neighbor**: 19.47% error rate
- **Bayesian Network**: 16.41% error rate
- **Neural networks**: 14.53% error rate

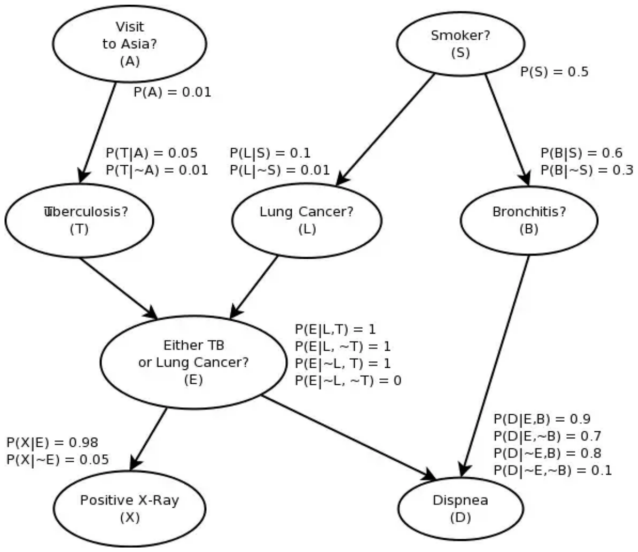
Brief discussion:

The average error rates for the top two methods are not significantly different. Having said that, the Bayesian Network method and Neural networks method have a gap of 2% in terms of classification precision.

Neural Networks is based on a large collection of neural units loosely modeling the way a biological brain solves problems with large clusters of biological neurons connected by axons. Neural nets are highly structured networks bases, and have three kinds of layers - an input, an output, and so called hidden layers, which refer to any layers between the input and the output layers. Each node (also called a neuron) in the hidden and output layers has a classifier.



Bayesian networks, is a probabilistic directed acyclic graphical model, a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph.



A potential reason to pick artificial neural networks (ANN) over Bayesian networks is the correlations between input variables. Bayesian networks like Naive bayes assumes that all input variables are independent. If that assumption is not correct, then it can impact the accuracy of the Naive Bayes classifier. An ANN with appropriate network structure can handle the correlation/dependence between input variables.

The difference in average error rate between the top ranked method (ANN) and the lowest ranked method (C4.5) is 6.63%. On that note, this is a noticeable difference that shows how the C4.5 approach to create a decision tree is mostly non optimal.

While ANN works with multiple input layers and validate multiple classifiers, C4.5 generates a decision tree where each node splits the classes based on the gain of information. The attribute with the highest normalized information gain is used as the splitting criteria. However, information gain will not always find the most optimal attribute to make the most efficient classifier. Finally, that explains why ANN outputs a better classifier with better results.

Question 6

Done with Pedro Casas, and submitted by himself.