Bartłomiej Kowalewski
nr 145204, L14
Wtorek 15:10

Pytanie 1:

1. The whole task is to say if a student pass or not the subject. Check the website `https://archive.ics.uci.edu/ml/datasets/student+performance` and analyze the list of attributes. Which of them would you consider as the most significant according to you? Choose 5-10 attributes, which will take part in the part of experiments.

Odpowiedź:

Subiektywnie wybrane najważniejsze atrybuty:

- **health** - current health status (numeric: from 1 - very bad to 5 - very good)
- **absences** - number of school absences (numeric: from 0 to 93)
- **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- **higher** - wants to take higher education (binary: yes or no)
- **internet** - Internet access at home (binary: yes or no)
- **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
- **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- **schoolsup** - extra educational support (binary: yes or no)

Pytanie 3:

3. Determine which metrics will be proper for the given datasets. Report three the most accurate metrics.

Odpowiedź:

- **accuracy**
- **mean absolute error**
- **root mean squared error**
- **confusion matrix**

Pytanie 4:

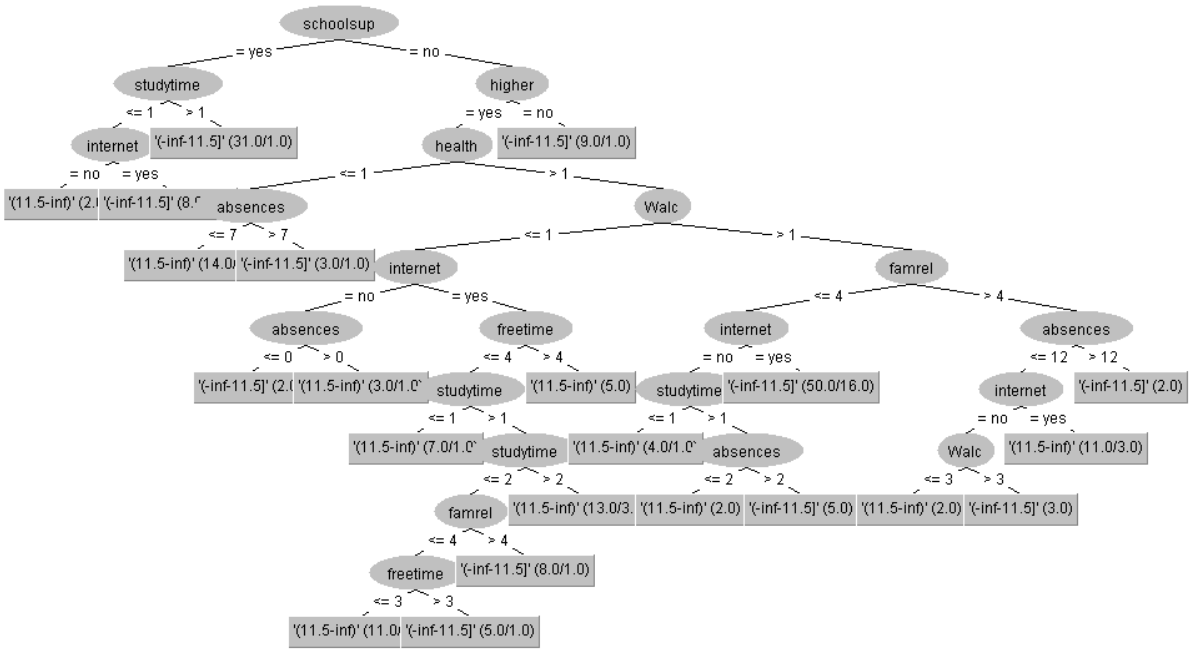
4. Get the chosen 5-10 attributes and test a few different values of the parameters, at least: confidenceFactor, minNumObj and binarySplits. Show the results for each set of parameters (you can visualize it also). For which set do you have the best result for test set?

Odpowiedź:

Instancja 1:

- confidenceFactor = 0.25
- minNumObj = 2
- binarySplits = False

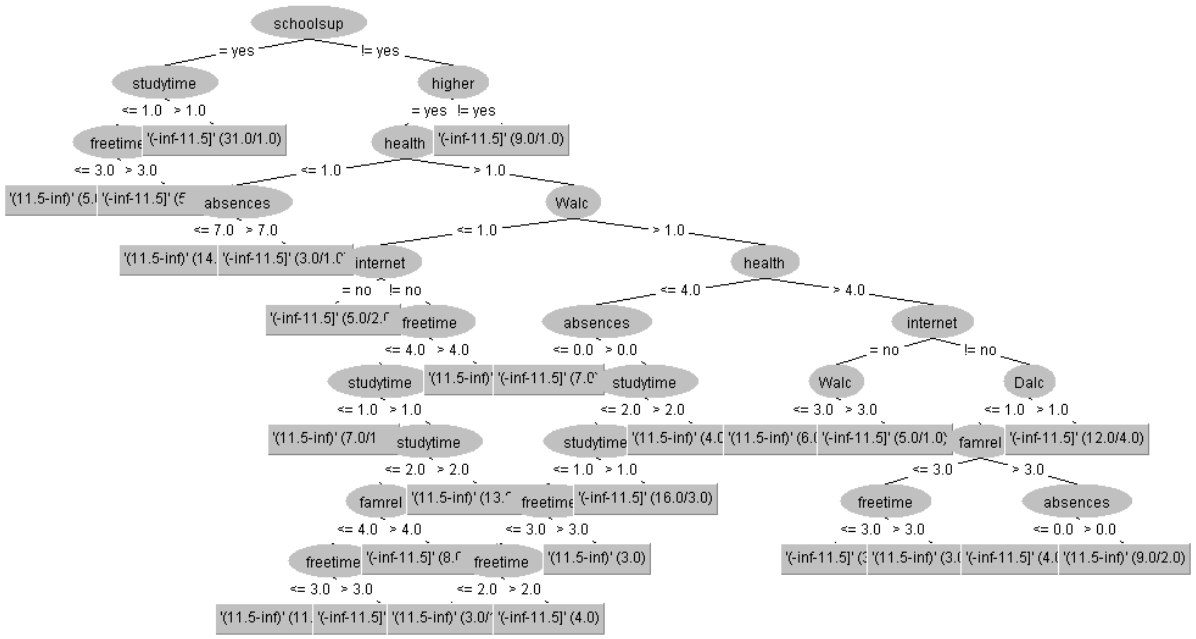
```
=== Summary ===

Correctly Classified Instances         109               55.8974 %
Incorrectly Classified Instances        86               44.1026 %
Kappa statistic                          0.0813
Mean absolute error                      0.4588
Root mean squared error                  0.5505
Relative absolute error                 94.7257 %
Root relative squared error            112.4037 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,627    0,545    0,638      0,627    0,632      0,081    0,567     0,649     '(-inf-11.
                 0,455    0,373    0,443      0,455    0,449      0,081    0,567     0,460     '(11.5-inf
Weighted Avg.    0,559    0,477    0,561      0,559    0,560      0,081    0,567     0,575

=== Confusion Matrix ===

  a  b   <-- classified as
 74 44 |  a = '(-inf-11.5]'
 42 35 |  b = '(11.5-inf)'
```



Instancja 2:

- confidenceFactor = 0.5
- minNumObj = 3
- binarySplits = True

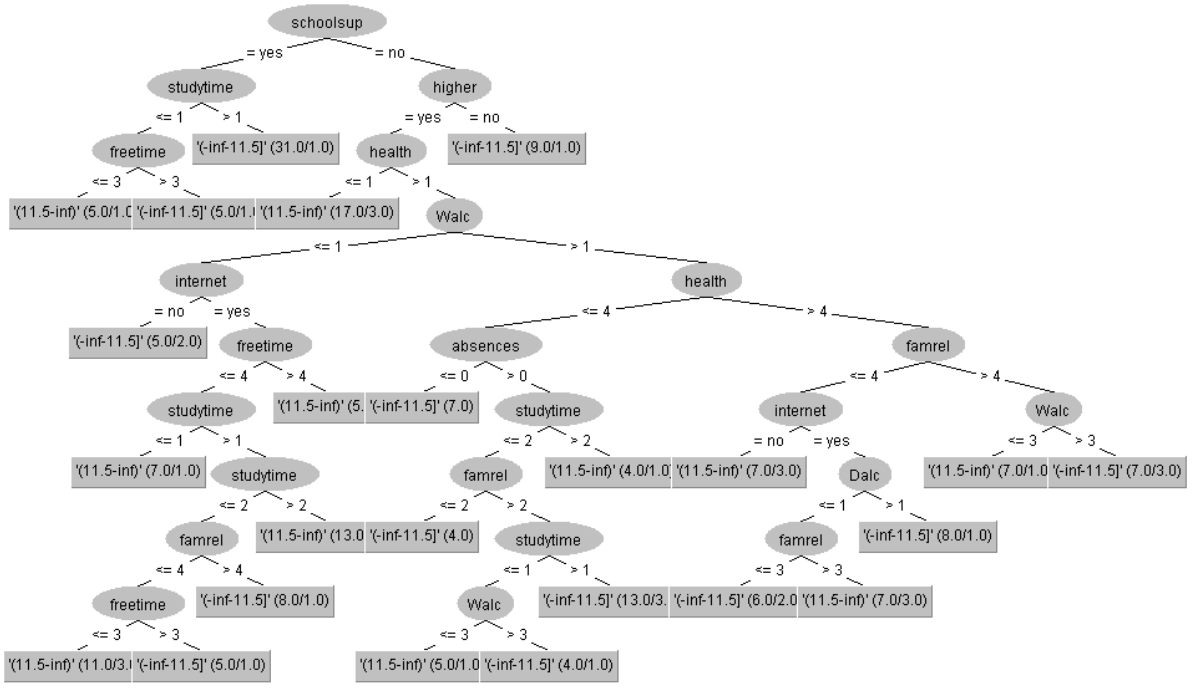
```
=== Summary ===

Correctly Classified Instances         110               56.4103 %
Incorrectly Classified Instances        85               43.5897 %
Kappa statistic                          0.094
Mean absolute error                      0.4622
Root mean squared error                  0.5647
Relative absolute error                 95.4204 %
Root relative squared error            115.288  %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,627    0,532    0,643      0,627    0,635      0,094    0,552     0,642     '(-inf-11.!
                 0,468    0,373    0,450      0,468    0,459      0,094    0,552     0,443     '(11.5-inf
Weighted Avg.    0,564    0,469    0,567      0,564    0,565      0,094    0,552     0,564

=== Confusion Matrix ===

  a  b   <-- classified as
 74 44 |  a = '(-inf-11.5]'
 41 36 |  b = '(11.5-inf)'
```



Instancja 3:

- confidenceFactor = 0.75
- minNumObj = 4
- binarySplits = False

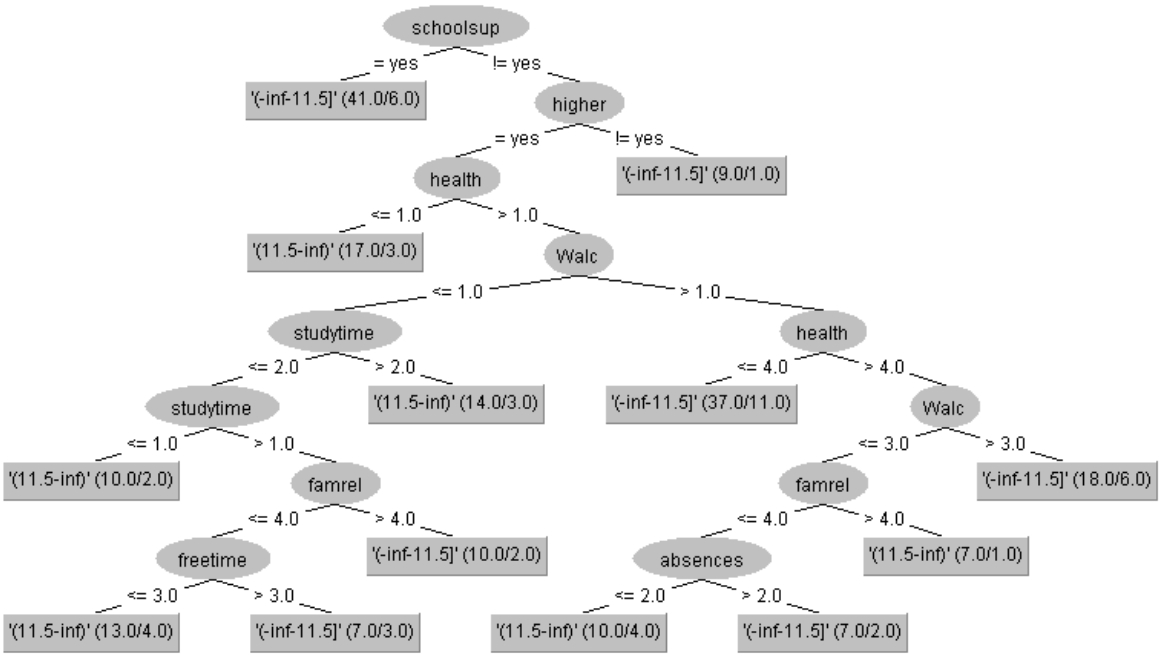
```
=== Summary ===

Correctly Classified Instances         105               53.8462 %
Incorrectly Classified Instances        90               46.1538 %
Kappa statistic                          0.0638
Mean absolute error                      0.4664
Root mean squared error                  0.5655
Relative absolute error                 96.2849 %
Root relative squared error            115.4615 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,559     0,494     0,635       0,559   0,595      0,064    0,547     0,629     '(-inf-11.!
               0,506     0,441     0,429       0,506   0,464      0,064    0,547     0,452     '(11.5-inf
Weighted Avg.  0,538     0,473     0,553       0,538   0,543      0,064    0,547     0,559

=== Confusion Matrix ===

  a  b   <-- classified as
 66 52 |  a = '(-inf-11.5]'
 38 39 |  b = '(11.5-inf)'
```



Instancja 4:

- confidenceFactor = 0.9
- minNumObj = 7
- binarySplits = True

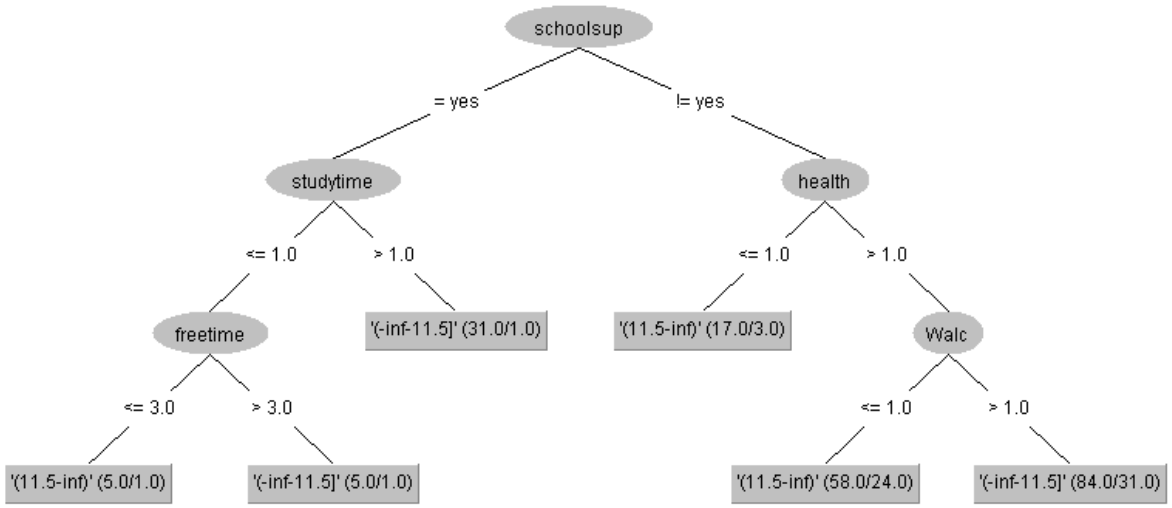
```
=== Summary ===

Correctly Classified Instances         110               56.4103 %
Incorrectly Classified Instances        85               43.5897 %
Kappa statistic                          0.0899
Mean absolute error                      0.4648
Root mean squared error                  0.5306
Relative absolute error                 95.9504 %
Root relative squared error            108.3356 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,636     0,545     0,641      0,636    0,638      0,090    0,574     0,651     '(-inf-11.5
               0,455     0,364     0,449      0,455    0,452      0,090    0,574     0,453     '(11.5-inf
Weighted Avg.  0,564     0,474     0,565      0,564    0,565      0,090    0,574     0,573

=== Confusion Matrix ===

  a  b   <-- classified as
 75 43 |  a = '(-inf-11.5]'
 42 35 |  b = '(11.5-inf)'
```



Instancja 5:

- confidenceFactor = 0.1
- minNumObj = 5
- binarySplits = True

```
=== Summary ===

Correctly Classified Instances         110               56.4103 %
Incorrectly Classified Instances        85               43.5897 %
Kappa statistic                          0.0817
Mean absolute error                      0.4769
Root mean squared error                  0.5125
Relative absolute error                 98.4546 %
Root relative squared error            104.638  %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,653    0,571    0,636      0,653    0,644      0,082    0,546     0,631     '(-inf-11.5
                 0,429    0,347    0,446      0,429    0,437      0,082    0,546     0,421     '(11.5-inf)
Weighted Avg.    0,564    0,483    0,561      0,564    0,563      0,082    0,546     0,548

=== Confusion Matrix ===

  a  b   <-- classified as
 77 41 |  a = '(-inf-11.5]'
 44 33 |  b = '(11.5-inf)'
```

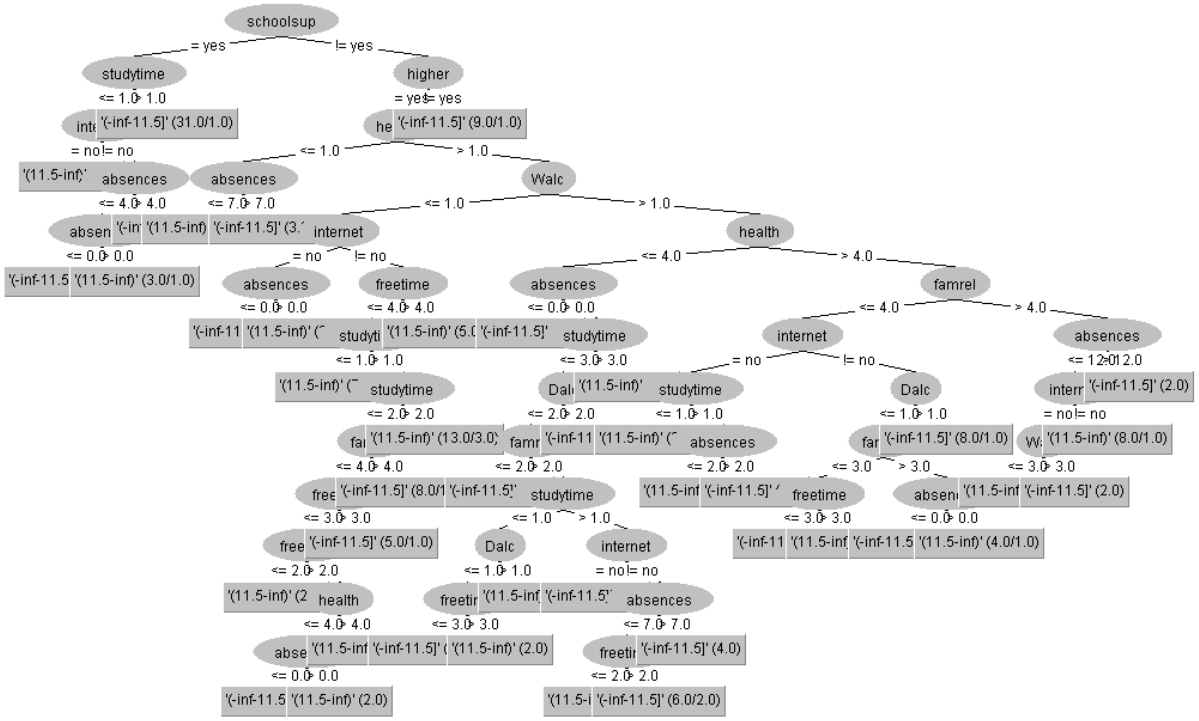


Instancja 6:

Dla tej instancji udało się osiągnąć najlepszy wynik.

- confidenceFactor = 0.9
- minNumObj = 2
- binarySplits = True

```
=== Summary ===

Correctly Classified Instances         116               59.4872 %
Incorrectly Classified Instances        79               40.5128 %
Kappa statistic                          0.1504
Mean absolute error                      0.4256
Root mean squared error                  0.5652
Relative absolute error                 87.8714 %
Root relative squared error            115.401  %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,669    0,519    0,664      0,669    0,667      0,150    0,580     0,644     '(-inf-11.
               0,481    0,331    0,487      0,481    0,484      0,150    0,580     0,487     '(11.5-inf
Weighted Avg.  0,595    0,445    0,594      0,595    0,594      0,150    0,580     0,582

=== Confusion Matrix ===

  a  b   <-- classified as
 79 39 |  a = '(-inf-11.5]'
 40 37 |  b = '(11.5-inf)'
```



Pytanie 5:

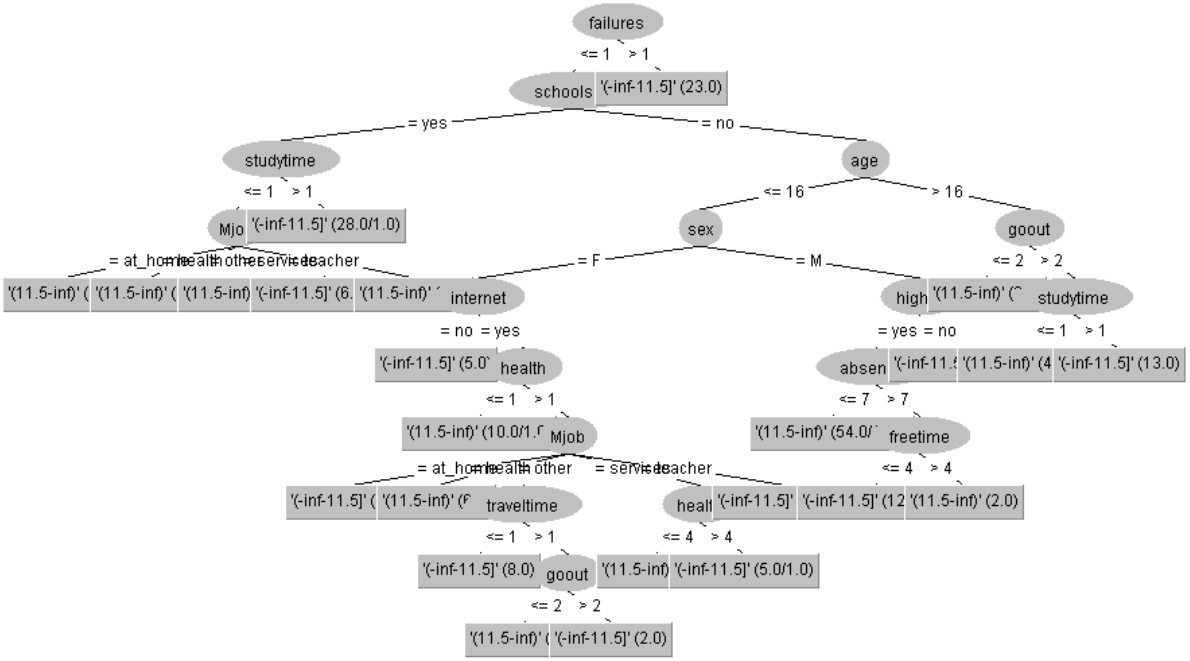
5. Repeat the previous step but this time get the whole set of attributes.

Odpowiedź:

Instancja 1:

- confidenceFactor = 0.25
- minNumObj = 2
- binarySplits = False

```
=== Summary ===

Correctly Classified Instances          99               50.7692 %
Incorrectly Classified Instances        96               49.2308 %
Kappa statistic                         -0.0302
Mean absolute error                      0.4798
Root mean squared error                  0.668
Relative absolute error                 99.0554 %
Root relative squared error            136.3731 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,593     0,623     0,593       0,593   0,593      -0,030   0,486     0,593     '(-inf-11.5
               0,377     0,407     0,377       0,377   0,377      -0,030   0,486     0,400     '(11.5-inf
Weighted Avg.  0,508     0,538     0,508       0,508   0,508      -0,030   0,486     0,517

=== Confusion Matrix ===

  a  b   <-- classified as
 70 48 |  a = '(-inf-11.5]'
 48 29 |  b = '(11.5-inf)'
```

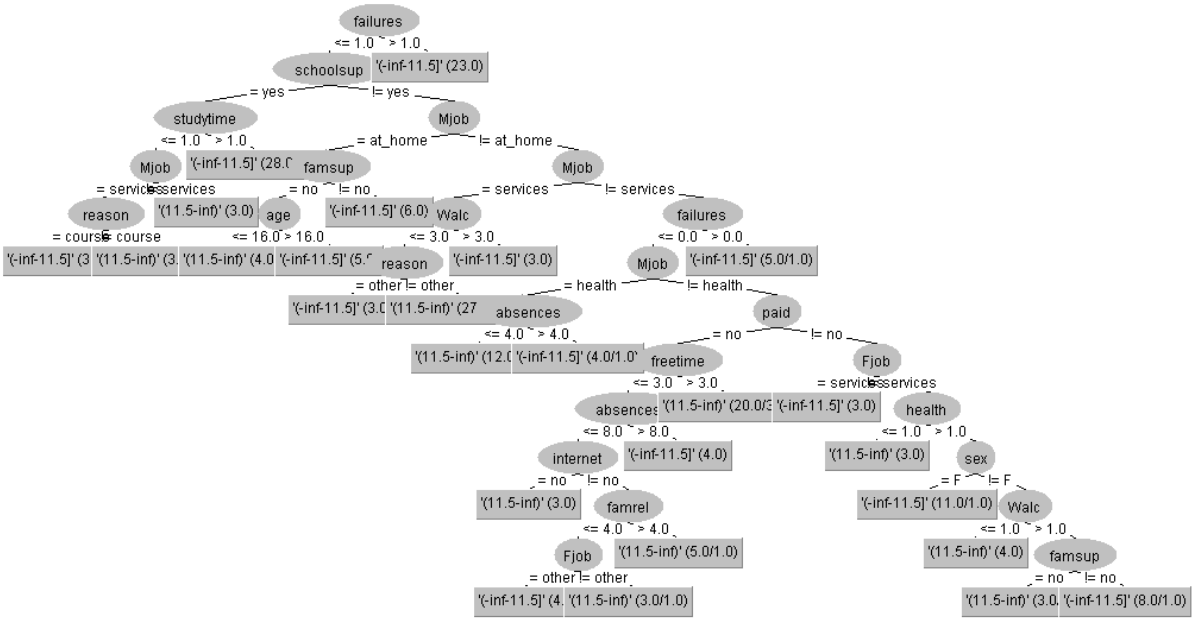


Instancja 2:

Dla tej instancji osiągnieto najlepszy wynik.

- confidenceFactor = 0.5
- minNumObj = 3
- binarySplits = True

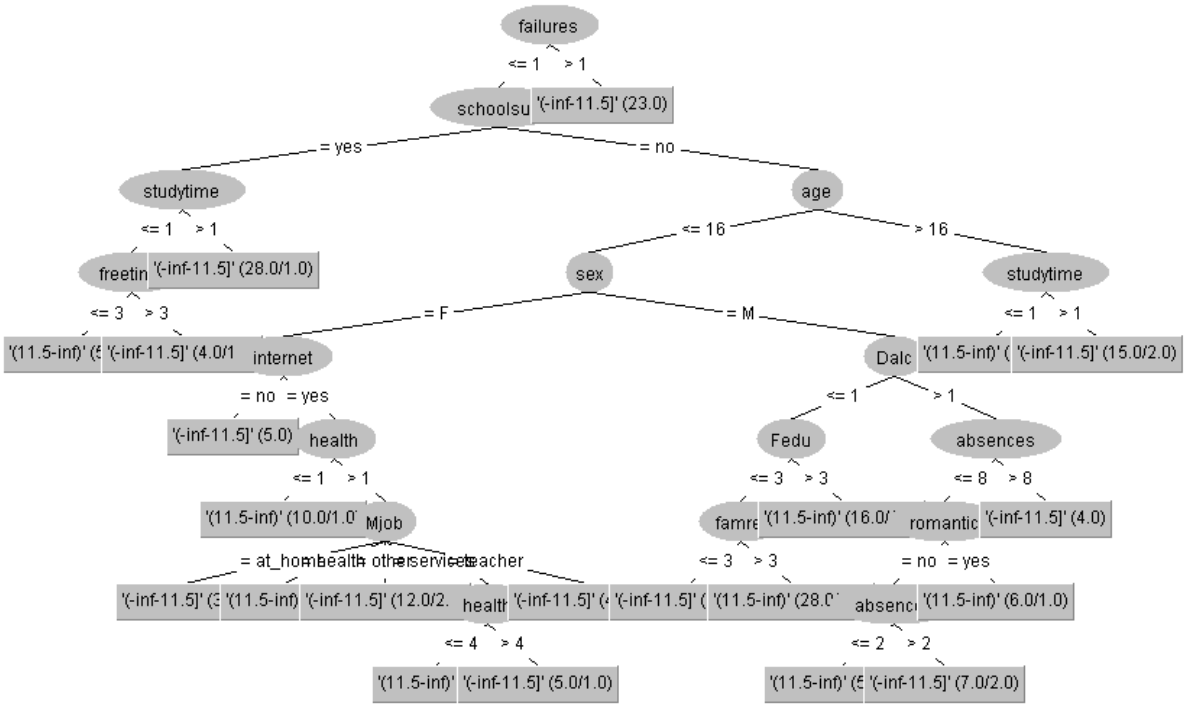
```
=== Summary ===

Correctly Classified Instances         126               64.6154 %
Incorrectly Classified Instances        69               35.3846 %
Kappa statistic                          0.2545
Mean absolute error                      0.3777
Root mean squared error                  0.5573
Relative absolute error                 77.9673 %
Root relative squared error            113.7903 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,720    0,468    0,702      0,720    0,711      0,255    0,613     0,668     '(-inf-11.
               0,532    0,280    0,554      0,532    0,543      0,255    0,613     0,498     '(11.5-inf
Weighted Avg.  0,646    0,393    0,644      0,646    0,645      0,255    0,613     0,601

=== Confusion Matrix ===

  a  b   <-- classified as
 85 33 |  a = '(-inf-11.5]'
 36 41 |  b = '(11.5-inf)'
```



Instancja 3:

- confidenceFactor = 0.75
- minNumObj = 4
- binarySplits = False

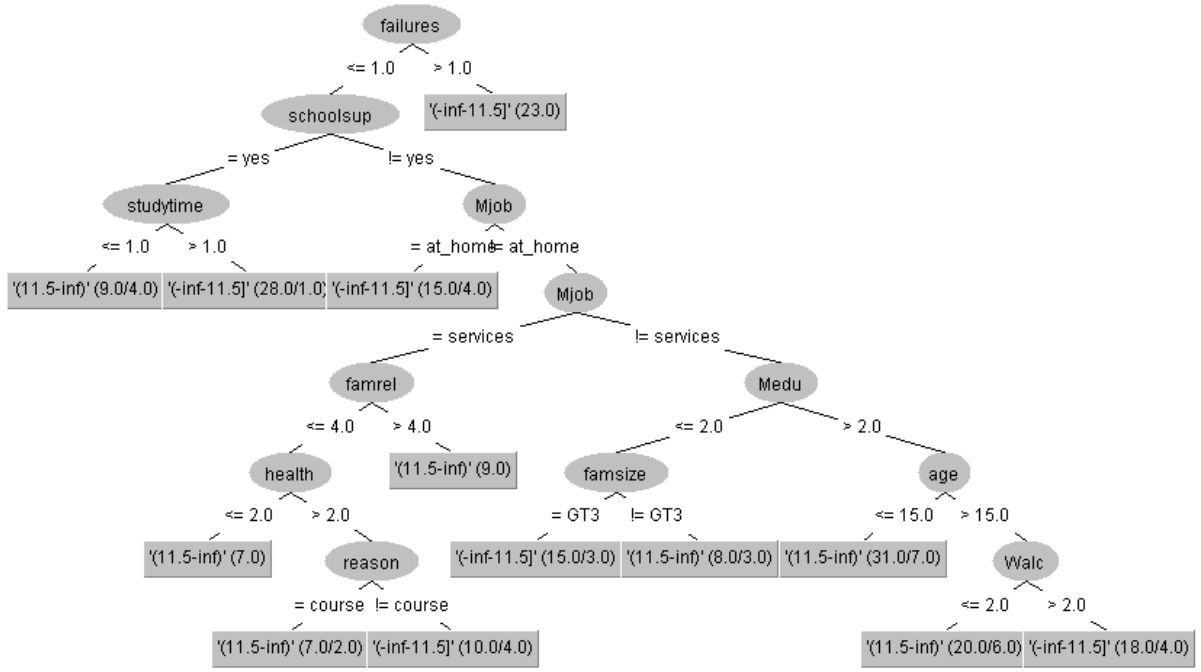
```
=== Summary ===

Correctly Classified Instances         104               53.3333 %
Incorrectly Classified Instances        91               46.6667 %
Kappa statistic                         -0.0602
Mean absolute error                      0.4621
Root mean squared error                  0.5902
Relative absolute error                 95.3996 %
Root relative squared error            120.5074 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,763    0,818    0,588      0,763    0,664      -0,066   0,507     0,622     '(-inf-11.5
                 0,182    0,237    0,333      0,182    0,235      -0,066   0,507     0,427     '(11.5-inf
Weighted Avg.    0,533    0,589    0,488      0,533    0,495      -0,066   0,507     0,545

=== Confusion Matrix ===

  a  b   <-- classified as
 90 28 |  a = '(-inf-11.5]'
 63 14 |  b = '(11.5-inf)'
```



Instancja 4:

- confidenceFactor = 0.9
- minNumObj = 7
- binarySplits = True

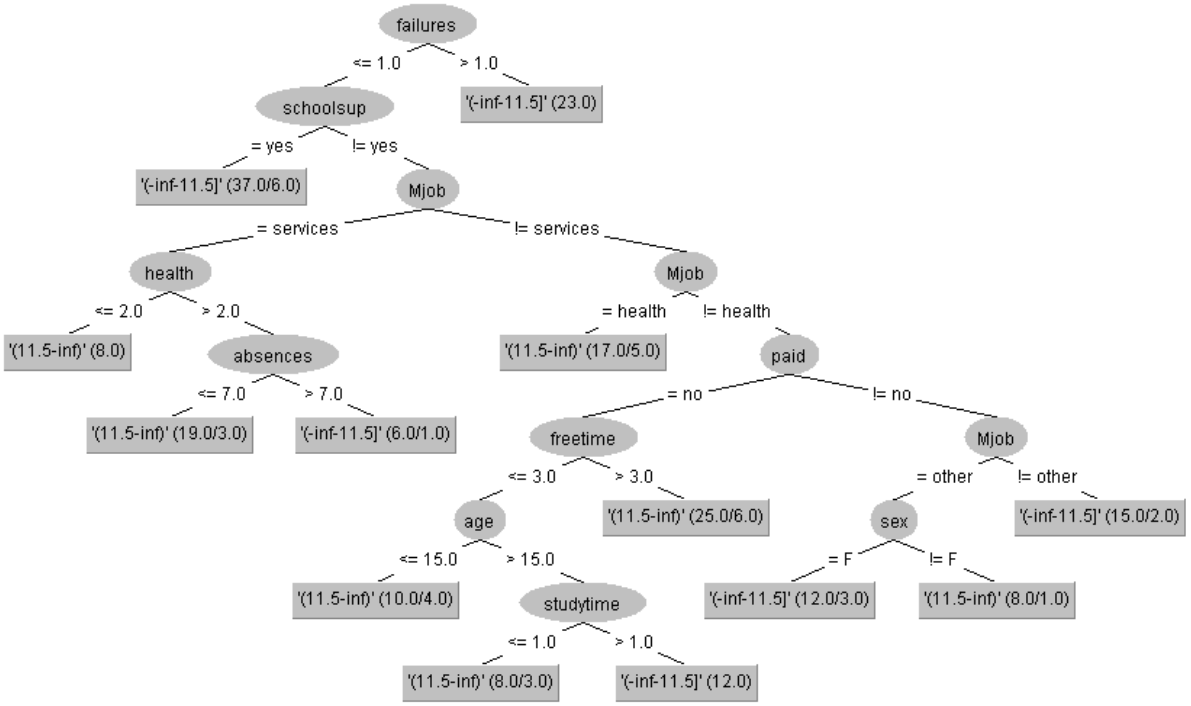
```
=== Summary ===

Correctly Classified Instances         111               56.9231 %
Incorrectly Classified Instances        84               43.0769 %
Kappa statistic                          0.0945
Mean absolute error                      0.4486
Root mean squared error                  0.5439
Relative absolute error                 92.6119 %
Root relative squared error            111.0397 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,653    0,558    0,642      0,653    0,647      0,095    0,575     0,663     '(-inf-11.5]'
               0,442    0,347    0,453      0,442    0,447      0,095    0,575     0,456     '(11.5-inf)'
Weighted Avg.  0,569    0,475    0,567      0,569    0,568      0,095    0,575     0,581

=== Confusion Matrix ===

  a  b   <-- classified as
 77 41 |  a = '(-inf-11.5]'
 43 34 |  b = '(11.5-inf)'
```



Instancja 5:

- confidenceFactor = 0.1
- minNumObj = 5
- binarySplits = True

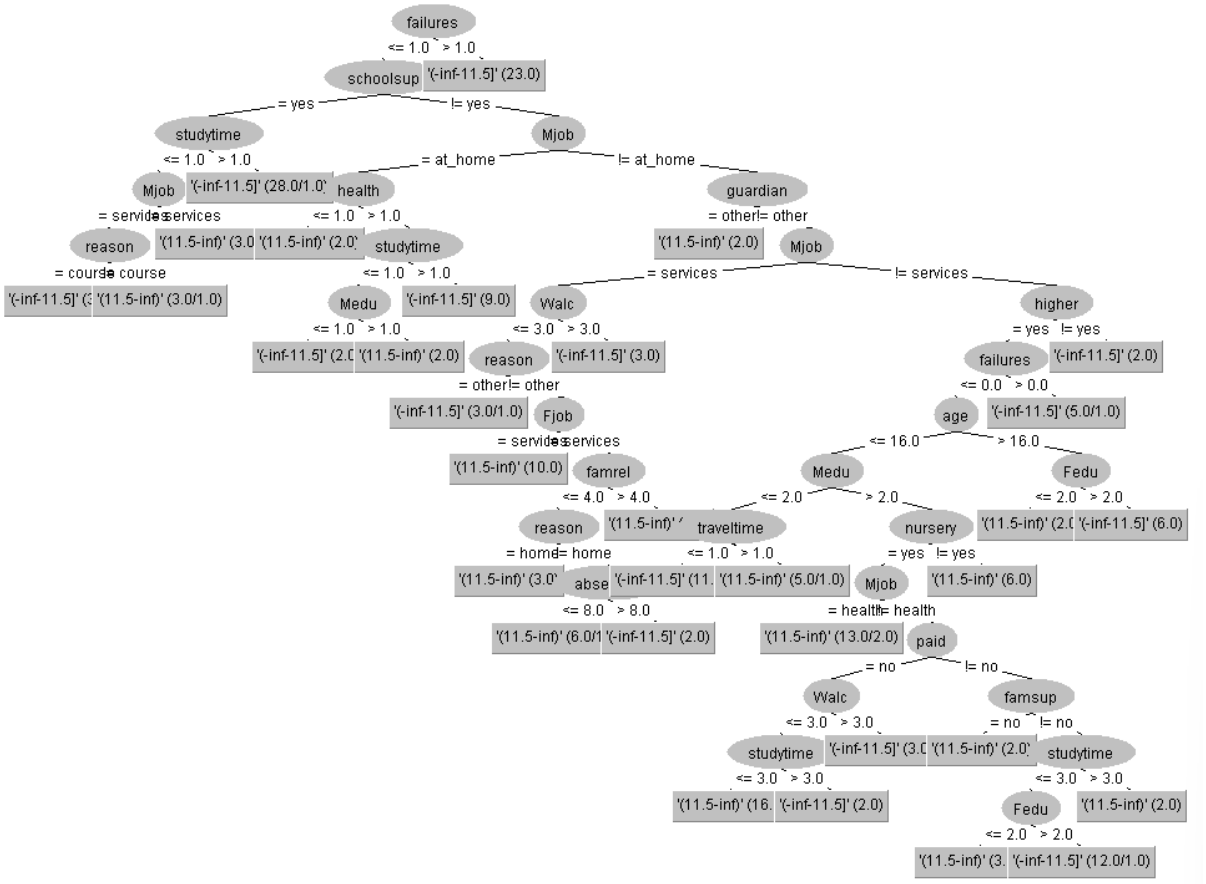
```
=== Summary ===

Correctly Classified Instances         107               54.8718 %
Incorrectly Classified Instances        88               45.1282 %
Kappa statistic                          0.0641
Mean absolute error                      0.4553
Root mean squared error                  0.5821
Relative absolute error                 93.9904 %
Root relative squared error            118.8501 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,610    0,545    0,632      0,610    0,621      0,064    0,543     0,642     '(-inf-11.5]'
               0,455    0,390    0,432      0,455    0,443      0,064    0,543     0,418     '(11.5-inf)'
Weighted Avg.  0,549    0,484    0,553      0,549    0,551      0,064    0,543     0,553

=== Confusion Matrix ===

  a  b   <-- classified as
 72 46 |  a = '(-inf-11.5]'
 42 35 |  b = '(11.5-inf)'
```



Instancja 6:

- confidenceFactor = 0.9
- minNumObj = 2
- binarySplits = True

```
=== Summary ===

Correctly Classified Instances         107               54.8718 %
Incorrectly Classified Instances        88               45.1282 %
Kappa statistic                          0.0683
Mean absolute error                      0.4631
Root mean squared error                  0.6658
Relative absolute error                 95.6028 %
Root relative squared error            135.9256 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,602     0,532     0,634       0,602   0,617      0,068    0,513     0,611     '(-inf-11.5]'
               0,468     0,398     0,434       0,468   0,450      0,068    0,513     0,405     '(11.5-inf)'
Weighted Avg.  0,549     0,479     0,555       0,549   0,551      0,068    0,513     0,530

=== Confusion Matrix ===

  a  b   <-- classified as
 71 47 |  a = '(-inf-11.5]'
 41 36 |  b = '(11.5-inf)'
```



Pytanie 6:

6. Load file *student-port* and run analysis for $k = 10$ in cross-validation. Present your results.

Odpowiedź:

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) G3_1

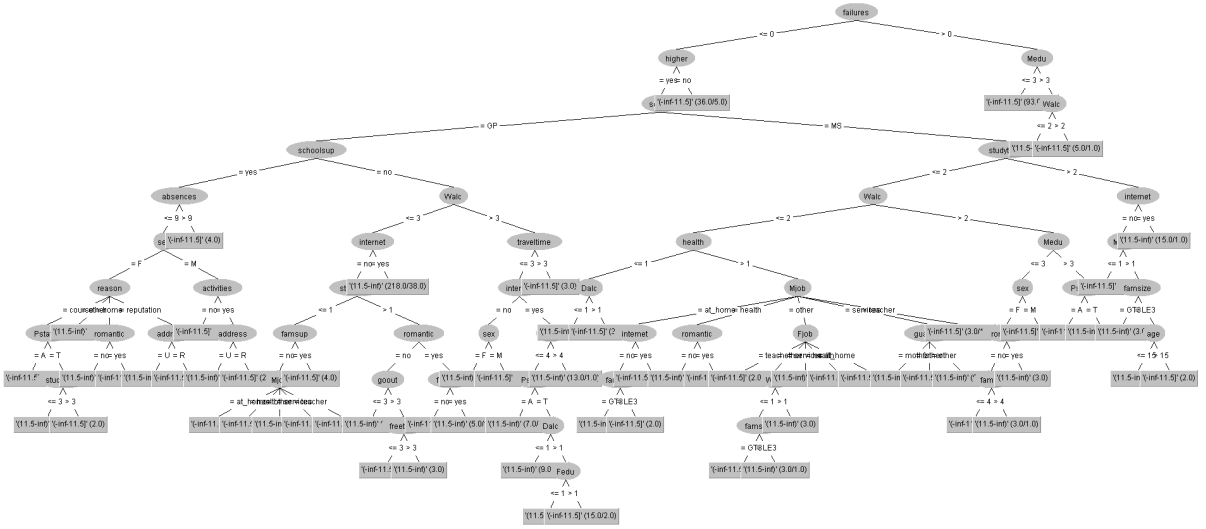
Start | Stop

Result list (right-click for options)

12:56:01 - trees.J48

Classifier output

```
Number of Leaves  :     65

Size of the tree :      117


Time taken to build model: 0.2 seconds

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

Correctly Classified Instances         432               66.5639 %
Incorrectly Classified Instances       217               33.4361 %
Kappa statistic                          0.3224
Mean absolute error                      0.3712
Root mean squared error                  0.5361
Relative absolute error                 74.6251 %
Root relative squared error            107.5087 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,581     0,261     0,658      0,581     0,617      0,324    0,633     0,566     '(-inf-11.5
               0,739     0,419     0,671      0,739     0,703      0,324    0,633     0,609     '(11.5-inf)
Weighted Avg.  0,666     0,346     0,665      0,666     0,663      0,324    0,633     0,589

=== Confusion Matrix ===

   a   b   <-- classified as
 175 126 |   a = '(-inf-11.5]'
  91 257 |   b = '(11.5-inf)'
```



Pytanie 7:

7. Both datasets (math and Portuguese) have the same set of attributes. Compare the structure of trees with the best results for each dataset. Are there any similarities between them? Basing on these results, can we say which attributes can say if the student is attentive?

Odpowiedź:

Do najważniejszych parametrów dla klasyfikacji możemy zaliczyć:

- **failures**
- **schoolsup**
- **studytime**

Wymienione parametry powtarzają się w drzewach skonstruowanych dla podanych instancji (zazwyczaj znajdują się w korzeniu lub bardzo blisko niego).

Pytanie 8:

8. Choose any other algorithm that you already know e.g. algorithm for rule induction that we used on first laboratories (PRISM) or Naive Bayes and run it on *student-mat* dataset. Compare the results from both algorithms. Which attributes had the biggest influence on the result? Are these attributes similar to those that you chose intuitively at the beginning of the task?

Odpowiedź:

```
=== Summary ===

Correctly Classified Instances         126               64.6154 %
Incorrectly Classified Instances        69               35.3846 %
Kappa statistic                          0.1961
Mean absolute error                      0.3623
Root mean squared error                  0.5179
Relative absolute error                 74.799  %
Root relative squared error            105.7407 %
Total Number of Instances              195

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,856    0,675    0,660      0,856    0,745      0,215    0,668     0,743     '(-inf-11.5]'
                 0,325    0,144    0,595      0,325    0,420      0,215    0,668     0,581     '(11.5-inf)'
Weighted Avg.    0,646    0,466    0,635      0,646    0,617      0,215    0,668     0,679

=== Confusion Matrix ===

   a   b   <-- classified as
 101  17 |   a = '(-inf-11.5]'
  52  25 |   b = '(11.5-inf)'
```

Wynik osiągnięty dla naiwnego klasyfikatora Bayesa jest praktycznie taki sam jak dla zbioru treningowego student-mat z wszystkimi atrybutami dla instancji:

- confidenceFactor = 0.5
- minNumObj = 3
- binarySplits = True

Atrybuty, które mają największy wpływ na rezultat powtarzają się z subiektywnie wybranymi przeze mnie cechami na początku zadania np. **studytime** i **schoolsup**.