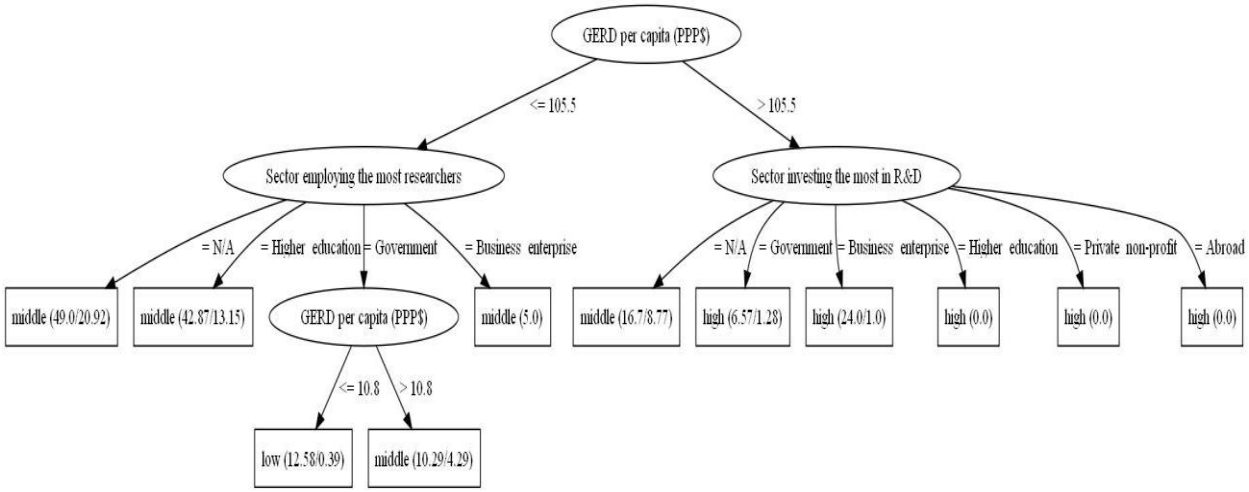
Questionnaire

General instructions: Please fill one page after another without looking ahead. Looking backwards is allowed, but not modifying your answers.
The data consists of 167 learning examples, each representing description of a country and a class (GNI per capita – low, middle, high). There are 27 attributes describing the R&D sector. The motivation is to find which attributes and relations contribute the most to the economic welfare of a country.
The first tree in Figure 1 was constructed in Weka. Please take a look at the tree, examine all the nodes in the tree and reply to the questions below. In the leaf there are two numbers. The first number denotes the number of examples in the leaf, and the second number represents the number of examples of non-majority classes.
While most of the attributes are comprehensible, GERD needs a short explanation: "GERD per capita" represents the level of investment in R&D in the relative form (PPP$ - purchasing power parity) to avoid direct link to the economic welfare, and "GERD as % of GNI" denotes percentage of GNI (gross national income), designated to research.
CCPE denotes corrected class probability estimate and is a measure showing how significant is the tree in comparison to all possible trees constructed on this data. 1 is the most significant and 0 the least. Accuracy and Kappa are measured in a 10-fold cross-validation.



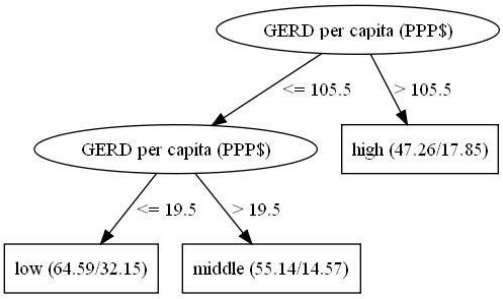
Accuracy : 64.67% Complexity : 11 relations Kappa: 0.4047 CCPE: 0.4113

Figure 1. The first tree.

Questions:

1. Does the tree sound reasonable or not?
   YES NO
2. Is the attribute in the root node (that the most important factor for the welfare of a country is the level of investment in R&D) reasonable or not?
   YES NO
3. Does the right subtree (starting with "Sector investing the most in R&D") present reasonable relations or not?
   YES NO
4. Does the left subtree (starting with "Sector employing the most researchers") present reasonable relations or not?
   YES NO

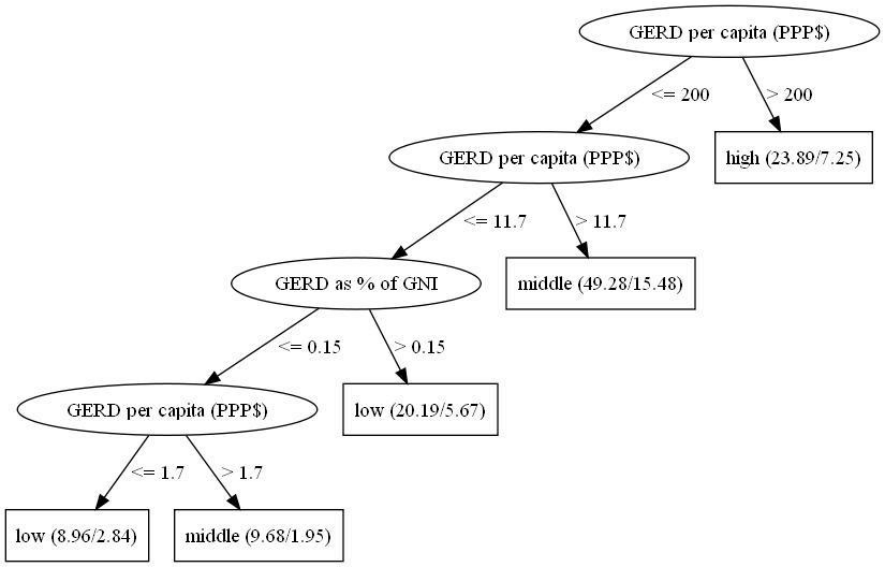
Three additional trees were generated with accuracy 63.47% (-1.2 percentage points compared to the tree in the previous page) and the other two with accuracy 66.47% (+1.8). Please take a look at the trees, examine all the nodes in the trees and reply to the questions on the next page. The questions from the first page are repeated, but this time additional information is provided.



Accuracy : 63.47% (-1.2) Complexity : 3 relations (-8)

Kappa: 0.4023 (-0.0024) CCPE: 0.5685 (+0.1572)

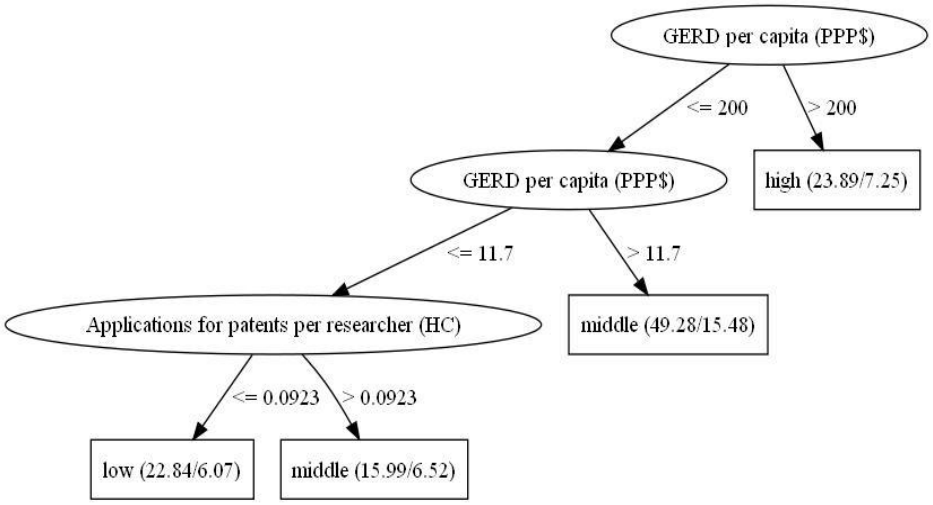
Figure 2. The second tree.



Accuracy : 66.47% (+1.8) Complexity : 5 relations (-6)

Kappa: 0.4314 (+0.0267) CCPE: 0.5287 (+0.1174)

Figure 3. The third tree.



Accuracy : 66.47% (+1.8) Complexity : 4 relations (-7)

Kappa: 0.4320 (+0.0273) CCPE: 0.5175 (+0.1062)

Figure 4. The fourth tree.

Questions:

5. Which of the additional three trees sounds more reasonable than the first one? *Please select one or more answers (e.g. c and d means that the third and the fourth tree sound more reasonable than the first one).*
    a. none
    b. second
    c. third
    d. fourth

6. What is your decision based upon? *Please select one answer.*
    a. accuracy
    b. complexity
    c. other measures: CCPE and/or Kappa
    d. content of the tree
    e. a combination of the above (e.g. b, c): _____

7. Write down the sequence of trees that persuaded you the most that the GERD attributes are the most important for the welfare of a country? E.g. 2, 3, 1, 4 means that 2 was the most persuasive and 4 the least.

    _____

8. Additional information to the question 3 from the first page: Five leaves in the right-hand side of the tree contain class “high” and only one “middle”. Analysis of the learning data reveals that there is actually no example with class “middle” in the data set corresponding to this subtree, but a couple of unknown values which the algorithm distributes among the default values cause this effect. Therefore, the whole right subtree could be represented by a leaf “high”. In light of this information please reply to the two questions.

   Do you agree with the explanation above?
   YES NO

   Does the right subtree (Sector investing the most in R&D) present a trust-worthy relation or not?
   YES NO

9. Additional information to the question 4 from the first page: Four leaves in the left-hand side of the tree contain class “middle” and only one “low”. Would not it be reasonable to substitute the node “Sector employing the most researchers” with a node below (GERD per capita), as in the second tree? In light of this information please reply to the next questions.

   Do you agree with the explanation above?
   YES NO

   Does the left subtree (Sector employing the most researchers) present trust-worthy relations or not?
   YES NO

   Would you prefer to substitute the left-hand subtree of the first tree with any of the left-hand subtrees of the third or the fourth tree?
   YES NO

   If the answer to the previous question was positive, please answer from which tree would you prefer to use the subtree? *Please select one answer.*
   THIRD FOURTH ANY