Further information on phase 8.1

Instructions for using the Jupyter Notebook <https://go.upb.de/auto-baum>

Pupils work in the learning environment in the form of a Jupyter Notebook, which can be used via "click and play". The use of the Jupyter Notebook is designed like a website and created as a learning environment so that pupils can largely work through it themselves. The learning environment includes the complete creation of a decision tree.

Ideally, the learning environment is worked through by pupils from top to bottom. Various activities are embedded, for which there is automatic feedback for the pupils.

Login name for the learning environment:

Password for the learning environment:

Before the activities/tasks are visible in the learning environment, all cells must first be executed once. This is done in the menu under Cell Run all

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

Automatisch generierte Beschreibung

The learning environment activities include the following elements.

* + **Data basis**  
    First, students are introduced to the structure of a data table and how it relates to the data cards. In a task, students are asked which value was transferred incorrectly from the data cards so that they can deal with both representations and their relationship.  
    *Solution: Food* apple *with the characteristic*
  + **Importing data**  
    The next step is to import the food data. This is done via the button Ein Bild, das Text, Schrift, weiß, Logo enthält.

    Automatisch generierte Beschreibung
  + **Create a training data set**  
    Students then assign labels for the training data set as in lesson 2. Here it makes sense to assign the same labels that were used in class. Either the set of cards with labels or worksheet 2 can be used for this.  
    Note: The labels can be changed again later!
  + **Sorting data**The data is sorted according to a selected characteristic in the next activity. At the touch of a button, the data is sorted in ascending order according to the values of the selected characteristic. Pupils recognize that this manual activity can be carried out much faster by the computer.
  + **Set up a decision rule**  
    As in the lesson, the next step is to find a threshold value. This is first done manually and can then be carried out automatically. The output contains the number of misclassifications.
  + **Create decision tree automatically**The computer then automatically creates a decision tree. The depth of the tree (as the number of levels) can be set so that pupils can gain experience with different numbers of levels of a tree. The maximum depth of the tree is 6, because features may not be reused at different levels.  
    Note: Different trees are created depending on the labels previously set for the food. Sometimes not all levels are required to generate the smallest number of misclassifications.
  + **Classify new foods**At the end of the learning environment, new foods can be classified by the tree.

Note on the source code in the Jupyter Notebook:

In the Jupyter Notebook, the students only work with menu-based elements. The source code is hidden. However, if the code of a cell is accidentally displayed in the Jupyter Notebook, "executing" this cell with the key combination Shift+Enter can hide the code again and the "beautiful" view can be restored.

Possibility of deepening:

* **Discussion of the human influence on the decision tree through data manipulation**

For in-depth analysis, the initial data in the Jupyter Notebook can be labeled differently (e.g. randomly, or all with the same label, etc.). This makes it possible to see what influence the data or the modeling of the data (here the human influence by assigning the labels) has on the results.

A decision tree is created for each training data set created in this way because the computer does not "know" whether its data basis is meaningful or not. However, if the training basis is "nonsensical", the automatically generated decision tree does not make sense either. This makes it clear that the quality of the data (modeling) is decisive for the quality of a rule system that is created using machine learning.