Teaching Decision Tree Using a Practical Example

Zdena Dobesova(✉)

Department of Geoinformatics, Faculty of Science, Palacky University,
17. listopadu 50, 779 00 Olomouc, Czech Republic
zdena.dobesova@upol.cz

Abstract. The positive experience with student gathering data in the Data Mining course is mentioned in this article. The inspiration for a practice lecture was a literature example of a decision tree for classification of sex from the weight and height of persons. Therefore, students fill anonymously simple questionnaire with personal weight, height, and sex. The data set was used as training data for the construction of a decision tree. Decision tree as supervised learning produces rules for classification of sex based on the input attributes. The final decision tree as a result of the training phase was used also for prediction of class (sex) on newly collected testing data. Both parts – construction of a decision tree and prediction was practically demonstrated. The data mining software Orange was selected for practical lectures. The Orange advantages are intuitiveness and easy design of workflow. The article shows result decision trees and results of prediction on real data. Teacher final finding is that the active collecting data make students more involved in the topic and assure a deep understanding of the lectured topics like decision trees.

Keywords: Data mining · Decision tree · Motivation · Engagement · Lecturing

1 Introduction

The author of the article guarantees the subject Data Mining in the first grade of master study program Geoinformatics at Palacký University in Olomouc, Czech Republic. The construction of a decision tree as a prediction model is one of the topics that belong to that course. The decision tree is a graphical expression of a set of rules that predict the final category. There are existing algorithms and their implementation in software for data mining and machine learning. The software automates the construction of classification trees. Students are familiar with two of the software for data mining tasks. The first is WEKA and the second is Orange [1, 2]. Both are non-commercial software for free use. The Orange is primarily aimed for educational purposes.

The practical examples help in the understanding of algorithms and software usage by university students. The author of the article has a positive experience with using a practical example to fix the knowledge of students in other courses. Contributive experience is mentioned in the article about the relational database design for the botanical garden plant database under the BotanGIS project [3].

R. Silhavy (Ed.): CSOC 2020, AISC 1226, pp. 247–256, 2020.
https://doi.org/10.1007/978-3-030-51974-2_23

2 Data and Methods

The important thing is the motivation of students to take part in some practical lessons. Acquiring knowledge is much stronger when students are engaged in gathering source data and when subsequently process this data as example data.

2.1 Decision Tree

The decision tree is a nonparametric algorithm of machine learning. The nonparametric method means that the learning process does not produce a form of function with learned coefficients like linear or logistic regression. Nonparametric algorithms are called machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data. Non-parametric methods are often more flexible, achieve better accuracy but require a lot more data and training time. Examples of include Support Vector Machines, Neural Networks and Decision Trees. Decision trees are an example of a low bias algorithm [4].

The graphical representation of the decision tree model is mostly a binary tree. Each node in a tree represents a single input variable and a split point on that variable. The leaf nodes of the tree contain an output variable (or target) which is used to make a prediction. Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node. Decision Trees are an important type of algorithm for predictive modeling machine learning [4]. Training data contains the output attribute – class value. Decision trees belong to supervised machine learning. They are also often accurate for a broad range of problems and do not require any special preparation for your data (numeric, categorical).

2.2 Literature Example as Inspiration

The literature contains some examples of decision trees. Most often is presented the dataset about playing tennis or golf (Yes/No) under various weather conditions (temperature, outlook, windy and humidity). The tennis-weather data is a small open data set with only 14 examples. In RapidMiner it is named Golf Dataset, whereas software WEKA has two data set: weather.nominal.arff and weather.numeric.arff [1, 5]. The resented final decision tree contains (Fig. 1) a first decision node (Outlook) with three branches (Sunny, Overcast and Rainy). Leaf node (Play) represents a classification or decision in the tree [6]. Figure 2 shows the decision tree based on training data set about playing tennis in WEKA software.

Book Master Machine Learning Algorithms [4] presents dataset with two input values of height in centimeters and weight in kilograms of some persons. The output value is sex as male or female. For demonstration purposes only, the binary classification decision tree is presented in that book (Fig. 3).

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Fig. 1. Example of data and corresponding decision tree for play golf [6].

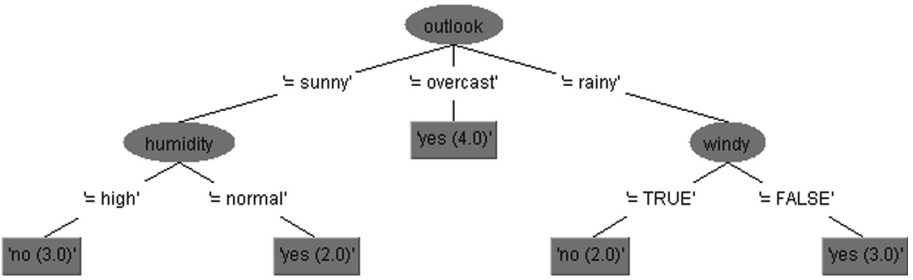


Fig. 2. Example of the decision tree for playing tennis in WEKA software [7].

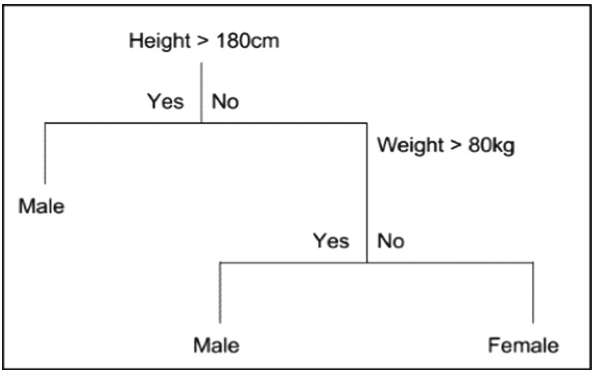


Fig. 3. Example decision tree of sex presented in the literature that is fictitious for demonstration purpose only [4].

The decision tree could be simply rewritten to the set of three rules [4]:

```
If  Height > 180 cm then Male
If  Height <= 180 cm  AND Weight > 80 kg then Male        (1)
If  Height <= 180 cm  AND Weight <= 80 kg then Female
```

This second literature example was an inspiration for practicing at university course lectured by me. The reason was that the example is understandable for students without any supplementary knowledge from a specific area. The next reason was that the gathering of training data was easy. I asked a group of students to fill the data anonymously to the questionnaire about their bodies.

3 Practical Example

The data set for practical example was collected for two years during lecturing course at university. The students were asked to fill the questionnaire. The questionnaire was prepared by Google Forms. It contains only three attributes: Weight, Height and Sex. No other information about the name or surname was collected. The group was relatively homogenous, and the age was from 20 to 26 years approximately. The collection of the data was assured at the beginning of the lecture about classification and regression decision trees. The link to the final collected data was handed on to the students. Students could freely download data and start to process them at the lecture. There was one interesting situation. One record was wrong due to mixing up weight and height by one respondent in the questionnaire. It was not hard to detect it and repair it. The correction of real data had also an educational effect. Real data very often contains mistakes, on the contrary, an official training data. The data are depicted in Table 1.

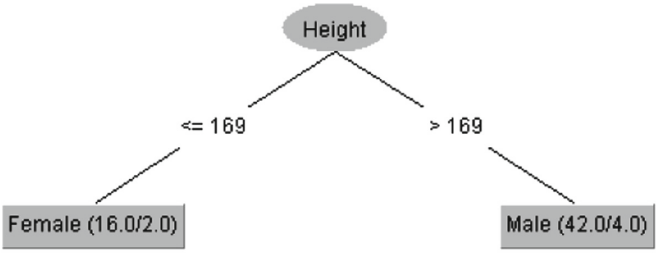
Table 1. Overview of gathered training data set for sex classification

Characteristic	Value
Number of records	58
Number of females	18
Number of males	40
Age	20–26 year
Weight: Min-Max	46–93 kg
Average weight	69.4 kg
Height: Min-Max	152–195 cm
Average height	176.3 cm

Firstly, students used the WEKA software. The WEKA implements J48 algorithm based on C4.5 [8] as an extension of the former ID3 algorithm [9, 10]. The constructed decision tree is in Fig. 4a. The condition for branching considers the value of 169 cm for person height. Totally six instances are incorrectly classified. The higher error is in the case of females where four females are classified as males. The WEKA also create a summary report with statistics (Fig. 4 b).

Secondly, students used the data mining software Orange. The algorithm for decision tree is designed in-house in Orange. There are three implementations

a)



b)

```
=== Summary ===

Correctly Classified Instances          52               89.6552 %
Incorrectly Classified Instances         6               10.3448 %
Kappa statistic                          0.7507
Mean absolute error                      0.1851
Root mean squared error                  0.3043
Relative absolute error                 43.0097 %
Root relative squared error             65.7591 %
Total Number of Instances               58

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,778    0,050    0,875      0,778    0,824      0,753    0,864     0,750     Female
                 0,950    0,222    0,905      0,950    0,927      0,753    0,864     0,894     Male
Weighted Avg.    0,897    0,169    0,896      0,897    0,895      0,753    0,864     0,849

=== Confusion Matrix ===

  a  b   <-- classified as
 14  4 |  a = Female
  2 38 |  b = Male
```

Fig. 4. Decision tree in WEKA software (a) and text summary (b).

(TreeLearner, SklTreeLearner, SimpleTreeLearner) [11]. The processing of data is designed like a workflow in Orange software (Fig. 5). The design is easy by drag and drop nodes to the canvas. The first node File (Men_Women dataset) at the left connects the source data. The setting of the decision tree is arranged by node Tree (in the middle). The dialogue offers the parameters like "Minimum number of instances in leaves", "Limit the maximal tree depth", etc. (Fig. 5). The parameters influence the pruning of a tree [12]. The last node on the right Tree Viewer visualizes the tree in the workflow.

The decision tree shows which attribute splits the best dataset. In the case of sex classification, it is the height (Fig. 6). The attribute height is more important that the attribute weight of a person in case the sex classification of persons. Only four instances are classified incorrectly. Orange uses the hue of color to depict the homogeneity of set in each node. The light and dark blue are for Female class in the presented tree. The light red and tones of red color are for Male class. Moreover, small circular diagrams on the right edges of rectangle nodes depict the structure of the

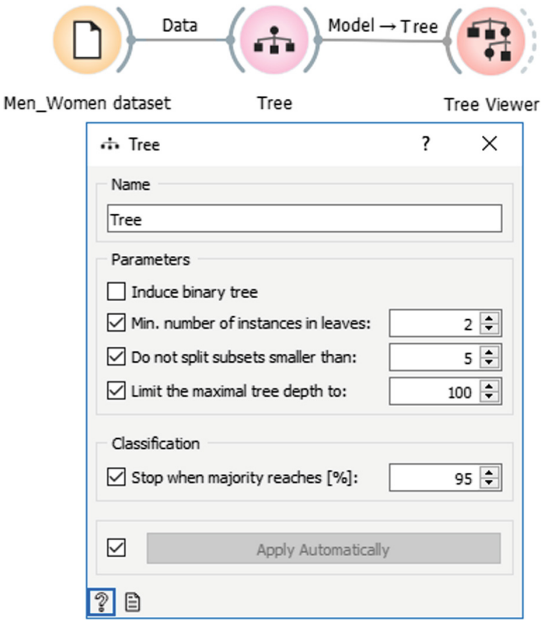


Fig. 5. Workflow for construction of a decision tree in Orange software with the dialogue of the node Tree.

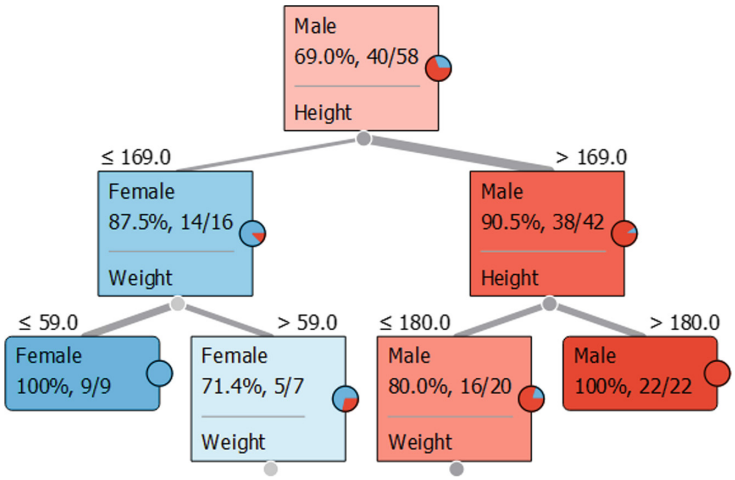


Fig. 6. Decision tree in Orange software for prediction of sex trained on 58 records.

dataset. The graphical expression of a tree is very illustrative and helps in the interpretation of a tree.

The final decision tree is possible to use for the prediction of new instances [13]. We used the newest dataset from the contemporary class of students as data for

prediction. The new dataset contains 13 persons (5 women and 7 men). The constructed workflow is the only an extension of the previous workflow (Fig. 5). New node File and Predictions are added to the workflow. Figure 7 shows the workflow and output table of node Predictions. The table contains all instances with the comparison of predicted sex and original data. The tree predicts eight instances correctly and five predicts incorrectly. Four females are predicted as males and one male is predicted as female. All women have a height higher than 169 cm in incorrectly classified instances.

It is evident that the prediction depends strongly on the training data for decision tree construction. The following step is trying to construct the decision tree using all data (totally 71 instances). The decision tree contains a top node ones again the weight, but the limit is 176 cm (Fig. 8). This practical experiment with different input dataset

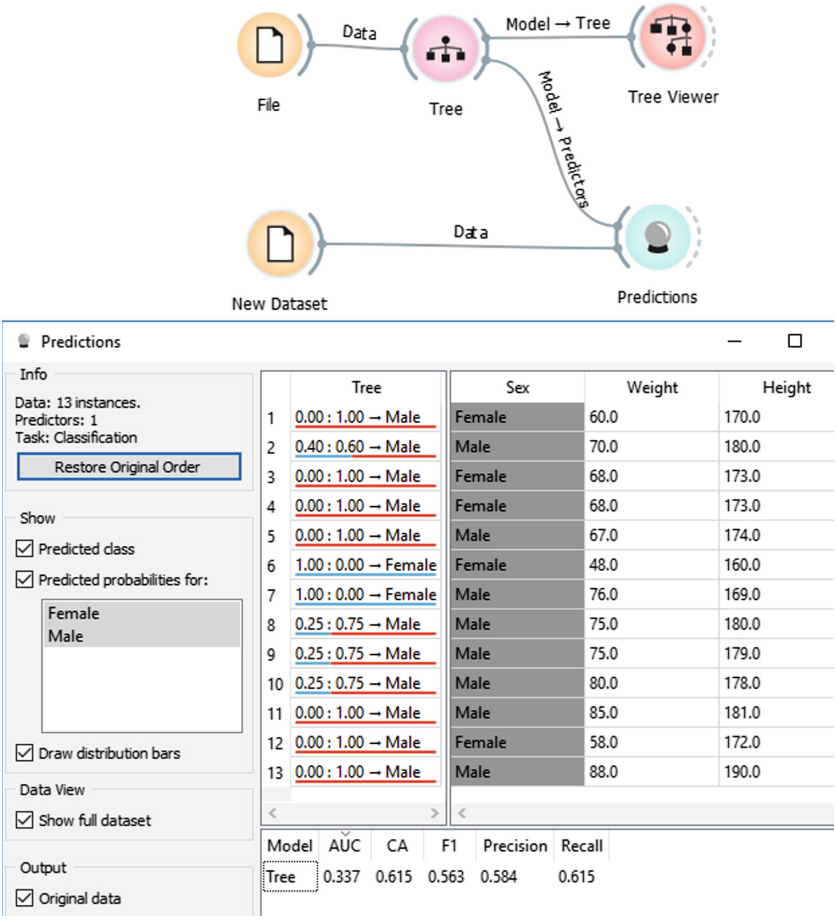


Fig. 7. Prediction of class sex in Orange software based on the pre-trained decision tree.

shows the influence of data and sensitivity on data values in practical lecture. Also, the situation of overfitting is preset.

There is space to compare and try the prediction with other methods like logistic regression [14]. The accuracy of the decision tree is 81% and logistics regression has 74%. In this case of all 71 instances, the logistic regression produces a worse result than the decision tree.

Beside this presented example also one more example is practised in lectures. The source data are also gathered locally at Palacky University. This data set is data about the dissemination of information about study branch Geoinformatics and geography between applicants for university study at secondary schools (the title is Dissemination of Study Information). The dissemination of information (leaflets, Geaudeamus fair-trade, Open Days, advertisements) is gathered by questionnaire for applicants [15]. The dataset has been systematically collected for four years, from 2016 to 2019.

The system of data collection and structure are presented at the one lecture of the Data Mining course like in the article [15]. Subsequently, the dataset is used two times in lecturing. Firstly, the construction of decision trees was used for practical prediction of the likelihood of high school students enrolling to study branch Geoinformatics and geography [16]. The second utilization is at the lecture about finding association rules [15] as another data mining method.

Both mentioned live data set (Sex Prediction and Dissemination of Study Information) are interesting for a practical demonstration of data processing in the course Data Mining. The data are more understandable for students than book examples and more attract them. On the other side, the interpretation of live data is more demandable because they may contain some outliers, errors or missing data.

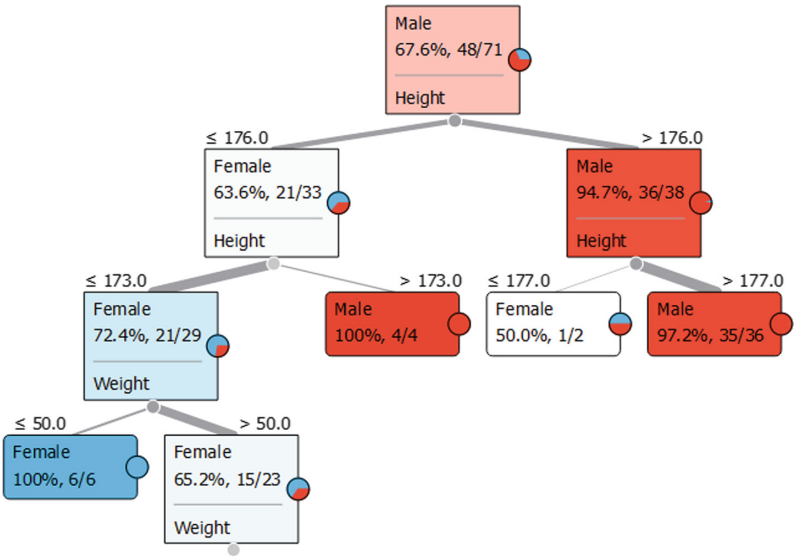


Fig. 8. Decision tree in orange for an extended data set of 71 instances.

4 Conclusion

The use of a practical example is promising. Students positively mentioned that example at written and oral exam while they finished the course. Therefore, this example will be used once again next year in the syllabus of subject Data Mining. Moreover, the data set could be easily extended by new records of new student personal measures. The model could be tested from the point of prediction and measuring model accuracy in case of new data records. The collected data also could be stored in the Moodle e-learning system for the course that runs at university. There is a potential to adapt them to one unified educational institution's datasets to serve for more courses and study branches like in other universities [17].

Moreover, there is a very good experience with Orange software for data mining tasks as educational software. It was the first experience with practicing Orange in the academic year 2019/20 at the Department of Geoinformatics as alternating for WEKA software. The use of Orange software is very intuitive and allows the quick design of workflow. The results of data mining are easily accessible for students in the process of acquiring new knowledge like using a decision tree method as one of the basic machine learning methods.

References

1. Machine Learning Group at the University of Waikato: WEKA, The workbench for machine learning. https://www.cs.waikato.ac.nz/ml/weka/. Accessed 15 Jan 2020
2. University of Ljubljana: Orange. https://orange.biolab.si/. Accessed 04 May 2019
3. Dobesova, Z.: Teaching database systems using a practical example. Earth Sci. Inf. **9**(2), 215–224 (2016)
4. Brownlee, J.: Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. Jason Brownlee (2016)
5. Nicôme - The Data Blog. https://gerardnico.com/data_mining/start. Accessed 10 Jan 2020
6. An Introduction to Data Science, Decision Tree - Classification. http://www.saedsayad.com/decision_tree.htm. Accessed 15 Jan 2020
7. Witten, I.H.: Data Mining. Morgan Kaufmann, Burlington (2009)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc. (1993)
9. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986)
10. Pavel, P.: Metody Data Miningu, Part 2. University of Pardubice, Economic-administrative faculty, Pardubice (2014)
11. University of Ljubljana: Orange Data Mining Library, Classification https://docs.biolab.si//3/data-mining-library/reference/classification.html#classification-tree. Accessed 16 Jan 2020
12. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston (2005)
13. University of Ljubljana: Getting Started with Orange 06: Making Predictions. https://www.youtube.com/watch?v=D6zd7m2aYqU. Accessed 10 Dec 2019
14. University of Ljubljana: Getting Started with Orange 07: Model Evaluation and Scoring. https://www.youtube.com/watch?v=pYXOF0jziGM. Accessed 16 Jan 2020

15. Dobesova, Z.: Discovering association rules of information dissemination about geoinformatics university study. In: Silhavy, R. (ed.) Artificial Intelligence and Algorithms in Intelligent Systems, vol. 764, pp. 1–10. Springer, Cham (2019)
16. Dobesova, Z., Pinos, J.: Using decision trees to predict the likelihood of high school students enrolling for university studies. Advances in Intelligent Systems and Computing, vol. 859, pp. 111–119 (2019)
17. Machova, R., Komarkova, J., Lnenicka, M.: Processing of Big Educational Data in the Cloud Using Apache Hadoop. IEEE (2016)