Module 4: Decision tree

14.06.2022

https://www.project-persist.eu

PERSIST.

PURCHASING EDUCATION RESEARCH SYNDICATE:
INDUSTRY 4.0 SKILLS TRANSFER

Co-funded by the Erasmus+ Programme of the European Union

UNIVERSITEIT TWENTE. technische universität dortmund UNIVERSITY OF ECONOMICS IN BRATISLAVA LUT University Edge Hill University

Decision trees and random forests

- ✓ Decision trees belong to supervised learning methods (subcategory of AI)
- ✓ Algorithms that classify data or predict outcomes accurately
- ✓ Supervised means that data columns have labels (both predicting and outcome variables)
- ✓ White-box solution for classification and regression problems:
  - ✓ Easy to understand
  - ✓ Logic can be visualized
  - ✓ Transparent

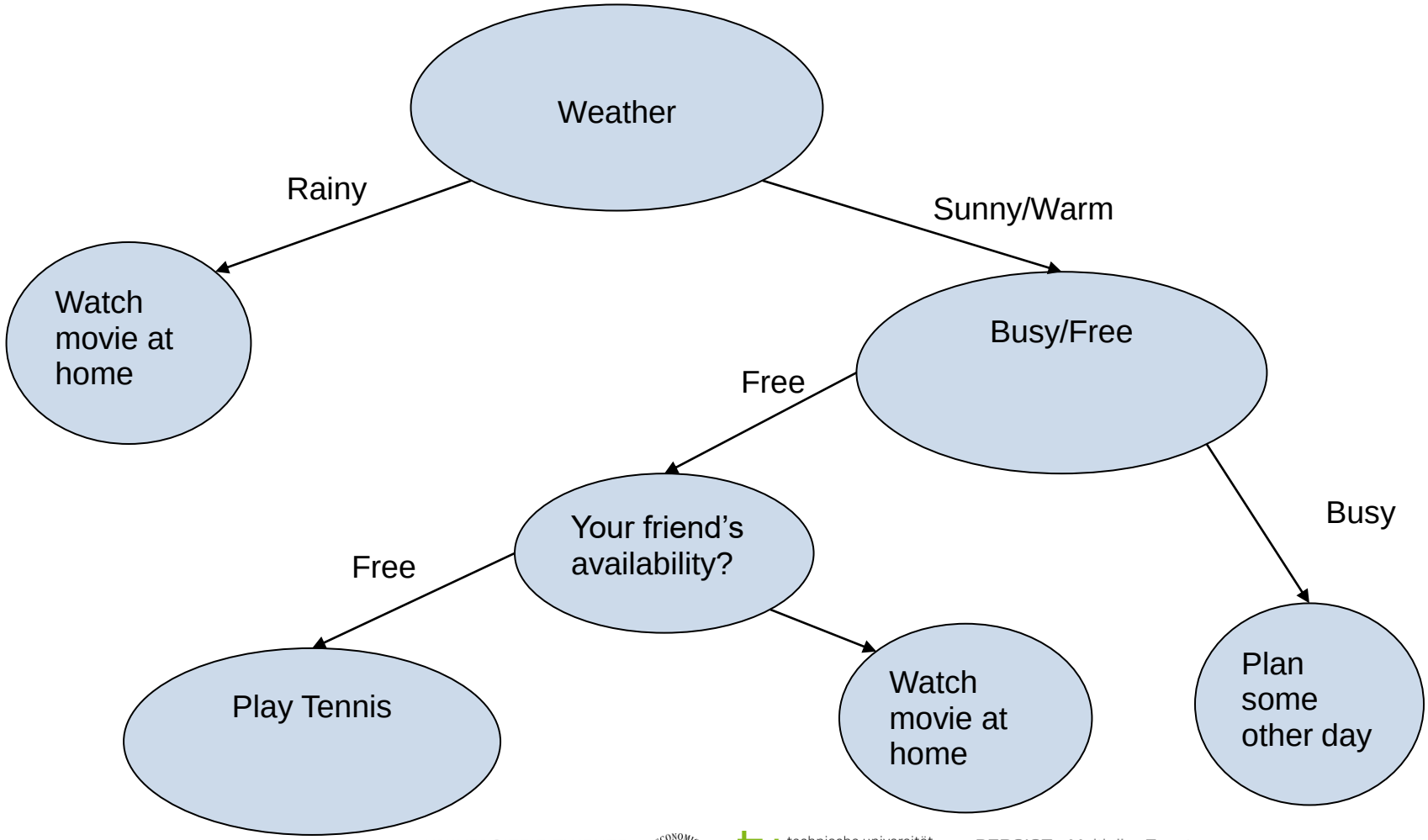
- ✓ Regression tree: These are used to predict continuous variables. For example, predicting rainfall in a region or predicting the revenue that a company might generate in the future.

- ✓ Classification tree: These are used to classify discrete variables. For example, classifying if the temperature of a day will be high or low, or predicting if a team will win the match or not

Decision tree limitations

1. Decision Tree Classifiers often tend to overfit the training data.
2. Changes in data may lead to unnecessary changes in the result.
3. Large trees can sometimes be very difficult to interpret.
4. These are biased toward splits on features having a number of levels.

Decision tree in Real-Life: Play Tennis



The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures(ASM) to split the records. (ASM e.g. entropy, information gain or Gini Index)
2. Make that attribute a decision node and breaks the dataset into smaller subsets.
3. Start building tree by repeating this process recursively for each child node until one of the conditions will match:
   - All nodes are close to pure
   - No more attributes to add

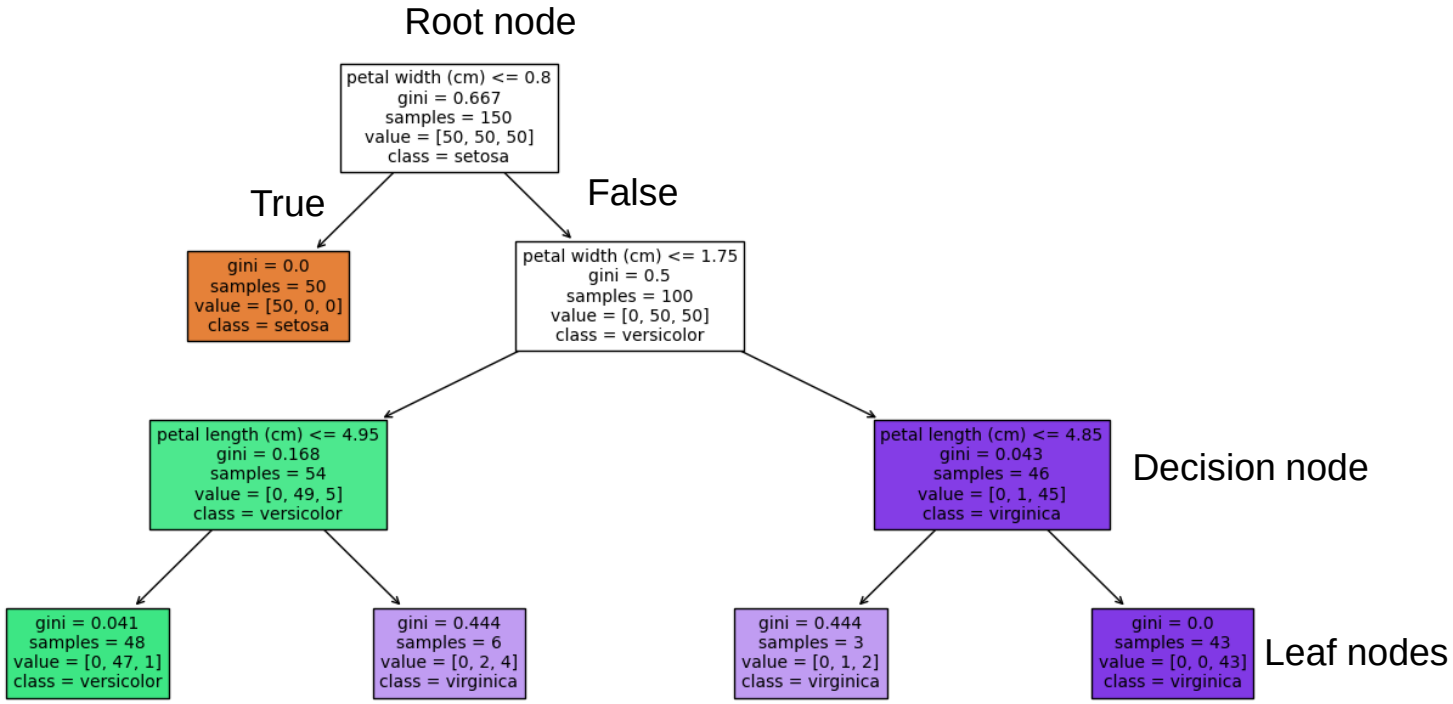
Objective: Find attributes which help to split the data into groups that are as pure as possible (i.e., homogeneous with respect to the target variable)

What it looks like in a Python script

- Many programming languages for machine learning
  - Python is one of them
- Libraries such as skicit-learn https://scikit-learn.org/stable/

Setosa, Versicolor and Virginica

Decision tree (here limited max depth = 3)



Entropy

Entropy measures the impurity or uncertainty present in the data.

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where:

- *S – set of all instances in the dataset*
- *N – number of distinct class values*
- *pi – event probability*

Entropy example

When entropy is smaller, the order is better in the tree

From the total of 14 instances we have:

- *9 instances "yes"*
- *5 instances "no"*

The Entropy is:

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

Information gain

The node that has a highest information gain, is the best root node

PERSIST.

How much information a particular variable or feature contributes to the model. Information gain basically tells the importance of a particular variable or feature toward the target variable or final result.

Information Gain (IG)

IG indicates how much "information" a particular feature/ variable gives us about the final outcome.

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j) = H(S) - H(A,S)$$

where:

H(S) – entropy of the whole dataset S

- ***|Sj| – number of instance with j value of an attribute A***
- ***|S| – total number of instances in dataset S***
- ***v – set of distinct values of an attribute A***
- ***H(Sj) – entropy of subset of instances for attribute A***
- ***H(A, S) – entropy of an attribute A***

Information gain example

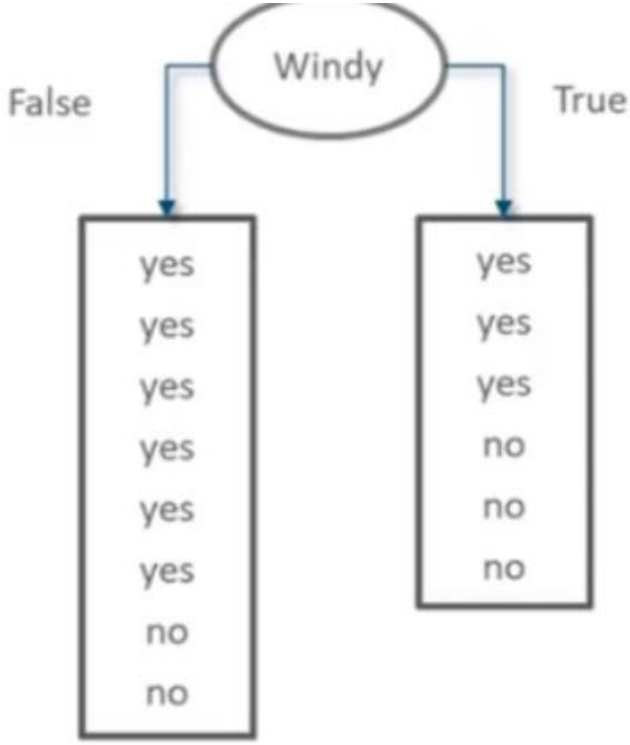
Information Gain of attribute "windy"

From the total of 14 instances we have:

- *6 instances "true"*
- *8 instances "false"*

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j)$$

$$Gain(A_{Windy}, S) = 0.940 - \frac{8}{14} \cdot \left(-\left(\frac{6}{8} \cdot \log_2 \frac{6}{8} + \frac{2}{8} \cdot \log_2 \frac{2}{8} \right) \right) + \frac{6}{14} \cdot \left(-\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6} \right) \right) = 0.048$$



Gini index

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Where, C is the total number of classes and p(*i*) is the probability of picking the data point with the class *i.*

If we have C=2 and p(1) = p(2) = 0.5, hence the Gini Index can be calculated as,

$$G = p(1) * (1-p(1)) + p(2) * (1-p(2))$$

$$= 0.5 * (1-0.5) + 0.5 * (1-0.5)$$

$$= 0.5$$

1. Open the map first to the new tab in order to see it during the quiz
2. After that open the quiz and use map to answer quiz

Decison tree is changing if the training data is changed!

Accuracy score

How well the decision tree predicts?

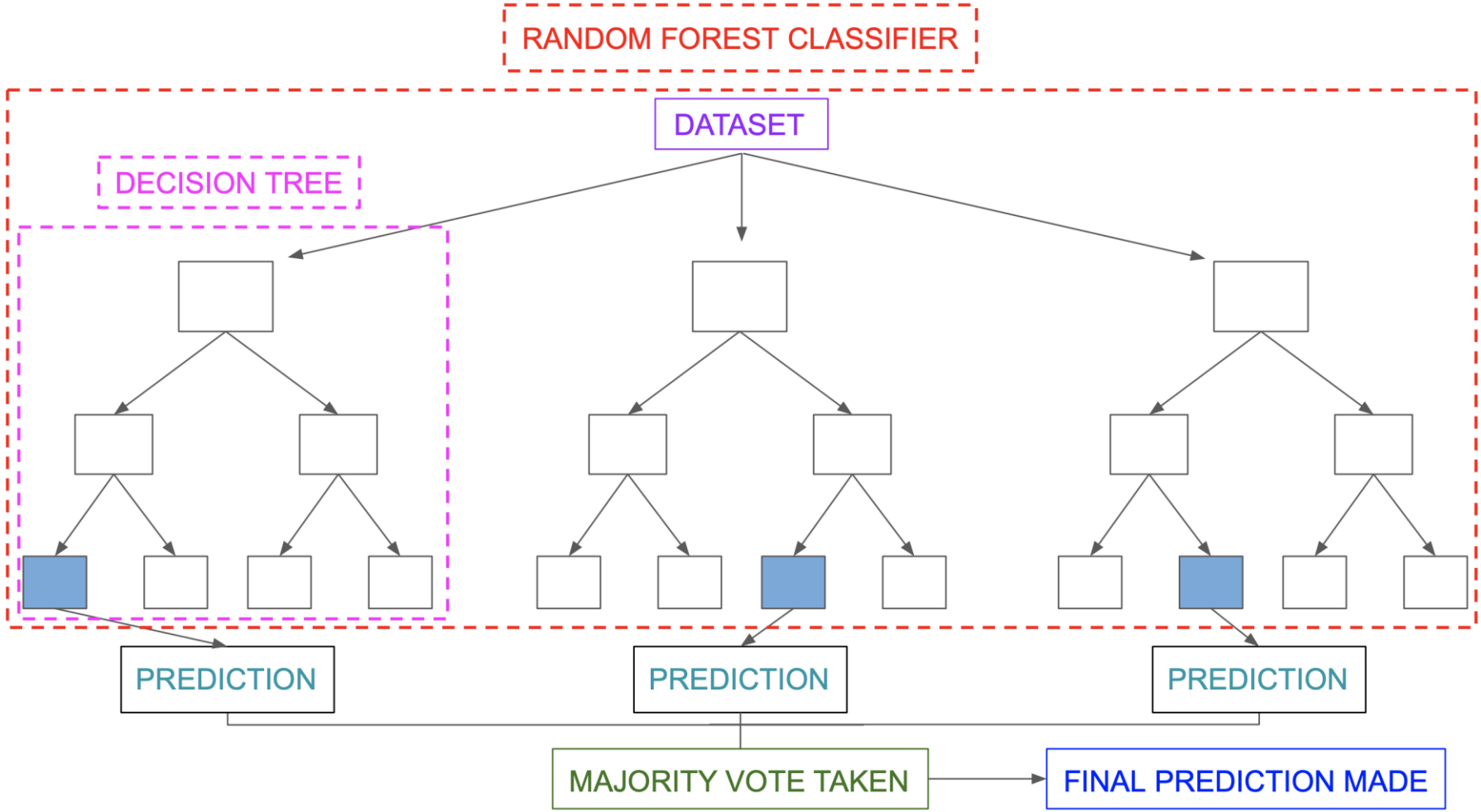
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Link in which decision trees are explained clearly (homework)

https://www.youtube.com/watch?v=ZVR2Way4nwQ

Random forests

Data for training is taken randomly from original dataset



Coffee break at 11.30 am, continue at 11.45 am, feedback somewhere at 12.15-12.30

10:45 – 12:30

Coffee break at 11.30 am,

continue at 11.45 am,
feedback somewhere at 12.15-12.30

The second part

-Start with Small wine test (10-15 min)

-Video: Data preparation for supervised machine learning (5 min)

-Video: Extra: Process automation: Twitter bot demo (14 min)

Small wine test in the moodle

30 min

Data of the exercise

1. https://en.wikipedia.org/wiki/Iris_flower_data_set
2. https://scikit-learn.org/stable/modules/tree.html

Source of random forests

https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

Contact Information

Website: http://project-persist.eu/
LinkedIn: PERSIST group
EMail: project-persist@utwente.nl

Next steps:

PERSIST.

PURCHASING EDUCATION RESEARCH SYNDICATE: INDUSTRY 4.0 SKILLS TRANSFER

Co-funded by the Erasmus+ Programme of the European Union

UNIVERSITEIT TWENTE. technische universität dortmund LUT University Edge Hill University

This research is part of Project PERSIST (project-persist.eu), co-funded by the Erasmus+ Programme of the European Union with project number 2019-1-NL01-KA203-060501.Disclaimer:TheEuropean Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, the Commission and the NA DAAD cannot be held responsible for any use which may be made of the information contained therein.