MachineLearnAthon - Microlecture Random Forest

Recorded by Lara Kuhlmann

MachineLearnAthon
A project Co-funded by the Erasmus+ programme of the European Union

Learning outcomes of today

After successfully completing this micro-lecture, you are able to….

- Explain the structure and basic components of Decision Trees.
- Describe the Random Forest algorithm as an ensemble method combining multiple decision trees.
- Understand how Random Forest improves prediction accuracy through different techniques.
- Differentiate between Random Forest and other methods.
- Evaluate the strengths and weaknesses of Random Forest.

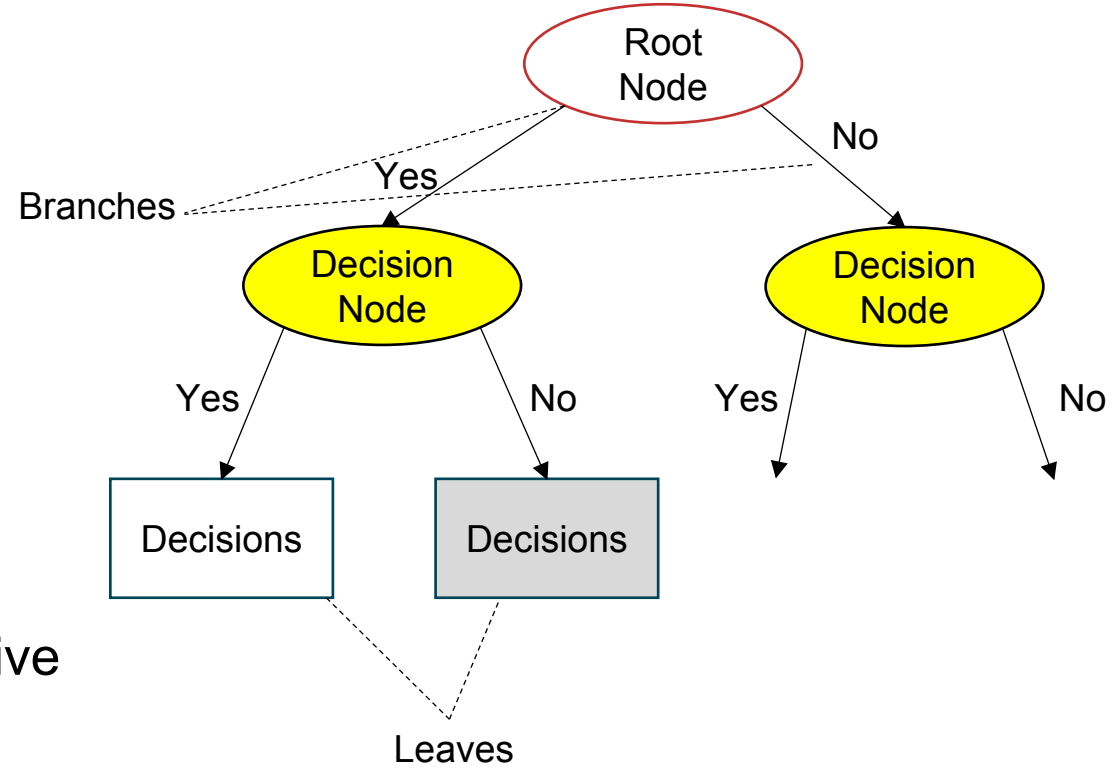
Co-funded by the Erasmus+ programme of the European Union

Agenda for today

- Definition of Decision Trees
- Definition of Random Forest
- Structure of Random Forest
- Functionality of the Random Forest
- Application example
- Strengths and weaknesses
- Comparison with other methods

Decision trees

- Models that structure data hierarchically through conditions.
- Structure:
  - Root node (Root Node)
  - Decision node (Decision Node)
  - Branches
  - Leaves
- Functionality: Determine root node, Splitting, recursive splitting, prediction
- Advantages: Intuitive results, no scaling necessary
- Disadvantages: Overfitting, instability, limited accuracy
  → Random Forest



Carl Kingsford & Steven L Salzberg (2008), S.1011 ff.

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

Random Forest

- Machine learning model developed by Leo Breiman and Adele Cutler in 2001.
  - Ensemble learning algorithm
  - Based on a large number of decision trees.
  - Combines results of several decision trees to make a prediction/ classification.
    - Each tree provides an individual prediction/ classification.
- Prediction for classification and regression problems

→ Decision by majority voting (for classification) or averaging (for regression).

- Often used in business and research

Steven J. Rigatti et al. (2017), S.31 ff. / Yanli Liu et al. (2012), S.246 / Davide Tramontin (2020), S.39

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

Functionality of the Random Forest

- Ensemble method: Combination of many decision trees for robust predictions.
- Bootstrap aggregation (bagging): Random samples with replacement

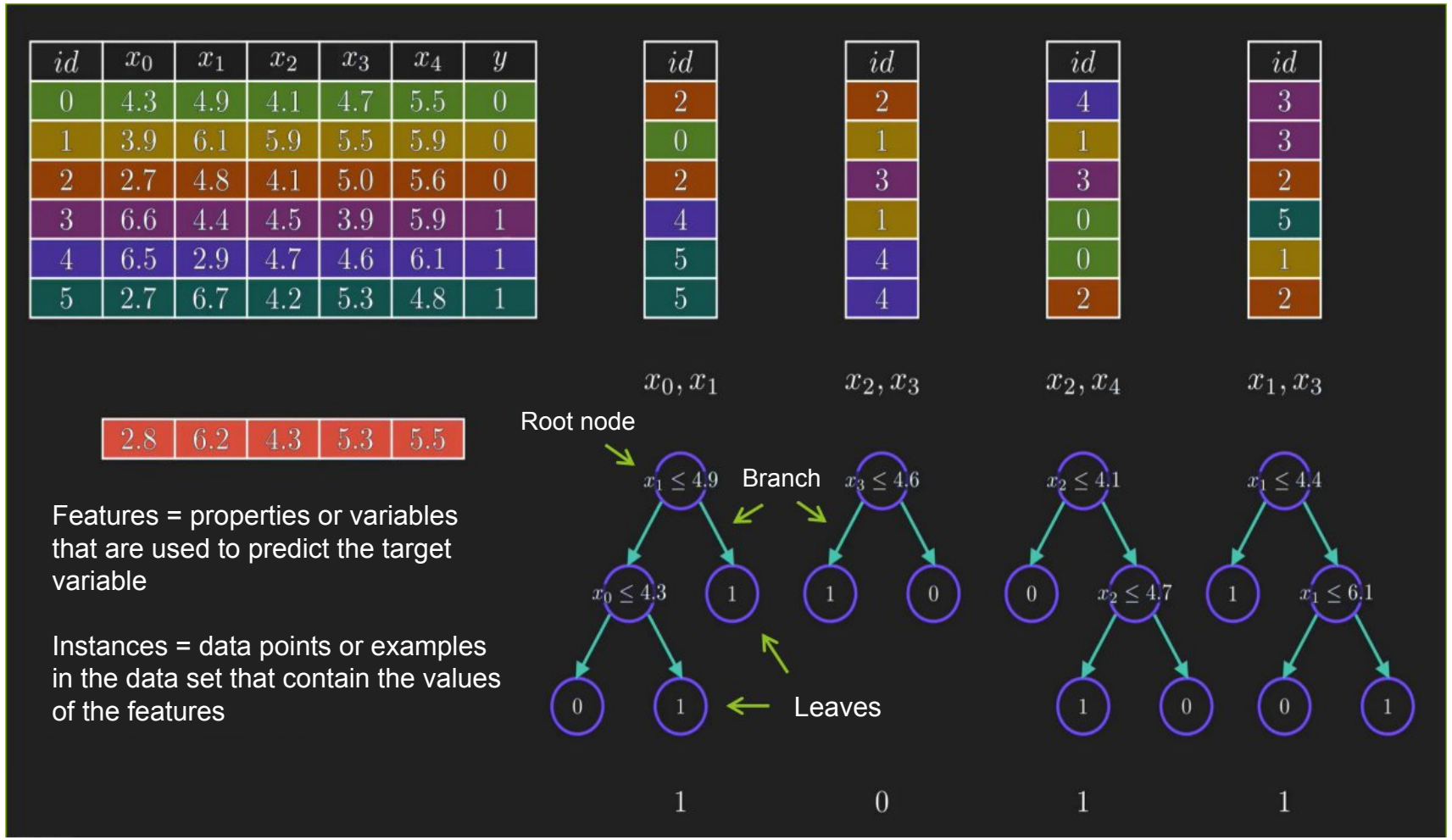
→ several subsets

- Random sampling ("bootstrap"): Ensure independent decision trees.
- Random feature selection: Each tree uses only limited features.
- Combined decisions: Majority vote or mean for prediction.
- More trees = Better accuracy: Irrelevant trees are ignored.

Steven J. Rigatti et al. (2017), S.32 f. / Leo Breiman (2001), S.5 ff.

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

Application example



- 5 features and 6 instances
- From training data
  Features and thresholds
  → Gini impurity, variance or information gain
- Each tree random subset of features and data points
  → Bootstrap sampling
- Majority vote and classification (y=0 or y=1)
- Out-of-bag data (OOB)
  Data points not used for training
- OOB data is used to evaluate the model performance

Random Forest Algorithm Clearly Explained! [Video]. YouTube. https://www.youtube.com/watch?v=v6VJ2RO66Ag

Strengths and weaknesses

Strengths	Weaknesses
+ Increased accuracy → combines the predictions of many decision trees.	- High memory requirements → Due to the parallel creation and storage of numerous decision trees.
+ Reduction of overfitting → Random selection of data and features reduces fitting to training data.	- Loss of interpretability → Shows important features, but makes relationships difficult to understand.
+ Robustness against noise → The ensemble compensates for the inaccuracies of individual trees.	- Higher computing requirements → requires more computing power compared to a single decision tree
+ Flexibility → Can be used for both classification and regression.	- Slow prediction time → The prediction time can increase with very large models.

Steven J. Rigatti et al. (2017), S.32 ff. / Leo Breiman (2001), S.5 ff

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Comparison with other methods

Random Forest vs Gradient Boosting (XGBoost, LightGBM):

Features	Random Forest	Gradient Boosting
Operating principle	Trees are independent, parallel training possible	Trees are dependent, sequential training necessary
Robustness	Insensitive to noise	Can be influenced by missing data
Training time	Fast (parallel training)	Slow (sequential training)
Areas of application	Moderate data set	High and complex data set
Accuracy	Precise	Very precise

- Random forest: Ideal for a lot of irrelevant data or when fast, robust predictions are required.
- Gradient boosting (XGBoost, LightGBM): Perfect for large data sets and clean data when maximum precision is crucial.

Steven J. Rigatti et al. (2017), S.32 ff. / Leo Breiman (2001), S.5 ff

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Recap this lecture

After successfully completing this lecture, you are able to….

- Explain the structure and basic components of Decision Trees.
- Describe the Random Forest algorithm as an ensemble method combining multiple decision trees.
- Understand how Random Forest improves prediction accuracy through different techniques.
- Differentiate between Random Forest and other methods.
- Evaluate the strengths and weaknesses of Random Forest.

Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

This material was created with the assistance of an AI language model.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon