RICE | D2K
Data to Knowledge Lab

Welcome to DSCI 101

Introduction to Data Science

Week 12 Recap

- Introduction to Machine Learning
  - Supervised, unsupervised and semi-supervised learning
  - Real world examples

- Supervised learning
  - Regression vs. classification
  - Deep learning and AI

- Your first supervised learning model
  - K-nearest neighbor

Week 13 Preview

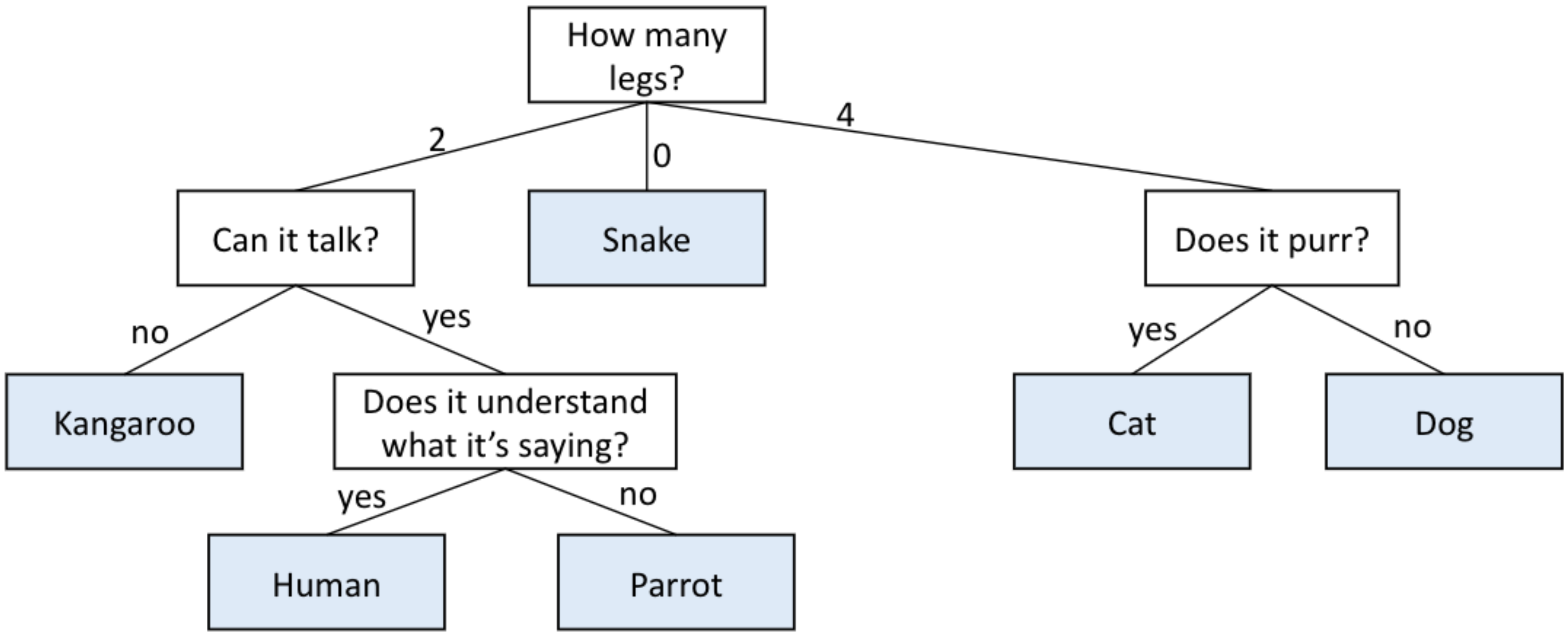
- Decision trees
  - for classification
  - for regression

- Advantages of tree models
  - interpretability
  - non-linear relationship with interactions

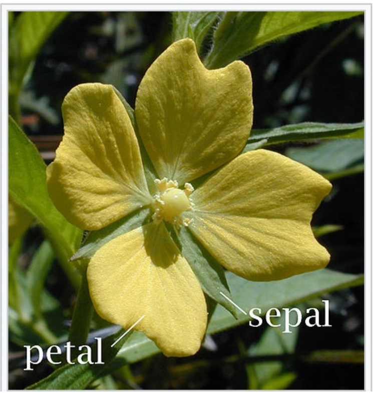
- Ensemble models
  - Random Forest

Decision Tree

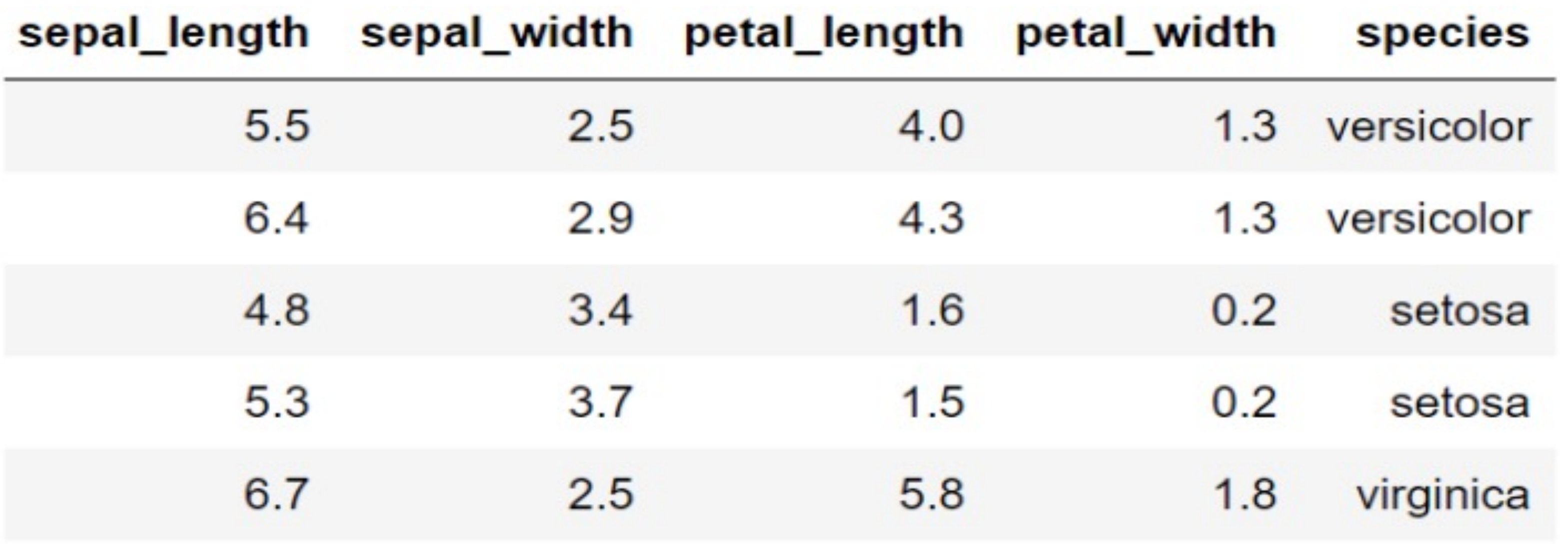
- A simple idea:



Example: Iris flower data set

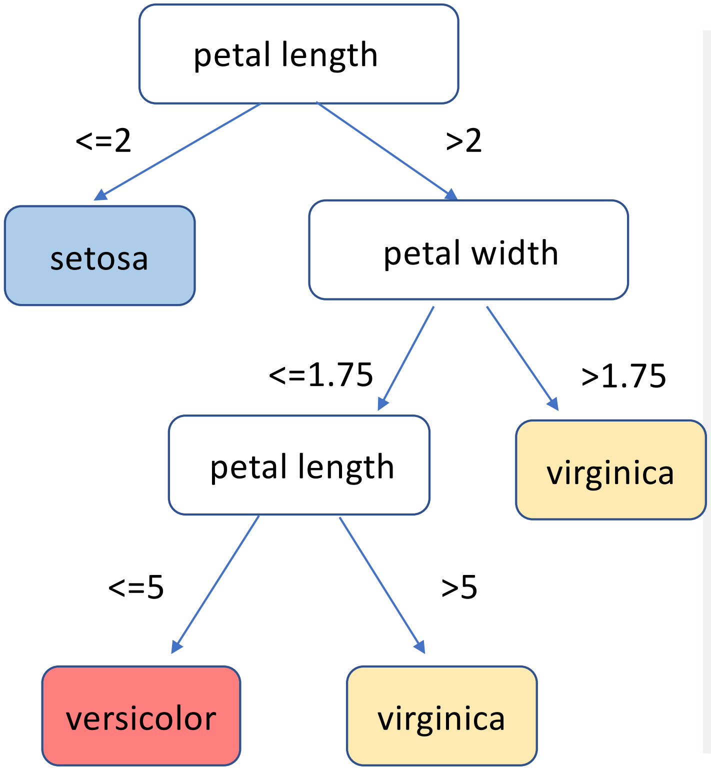


- Data set consists of 150 iris flower measurements
    - Columns: “petal length”, “petal width”, “sepal length”, “sepal width”, “species”
- Goal is to predict species from other columns / features
    - 3 different species



sepal_length	sepal_width	petal_length	petal_width	species
5.5	2.5	4.0	1.3	versicolor
6.4	2.9	4.3	1.3	versicolor
4.8	3.4	1.6	0.2	setosa
5.3	3.7	1.5	0.2	setosa
6.7	2.5	5.8	1.8	virginica

Tree for Classification



petal length

<=2

>2

setosa

petal width

<=1.75

>1.75

petal length

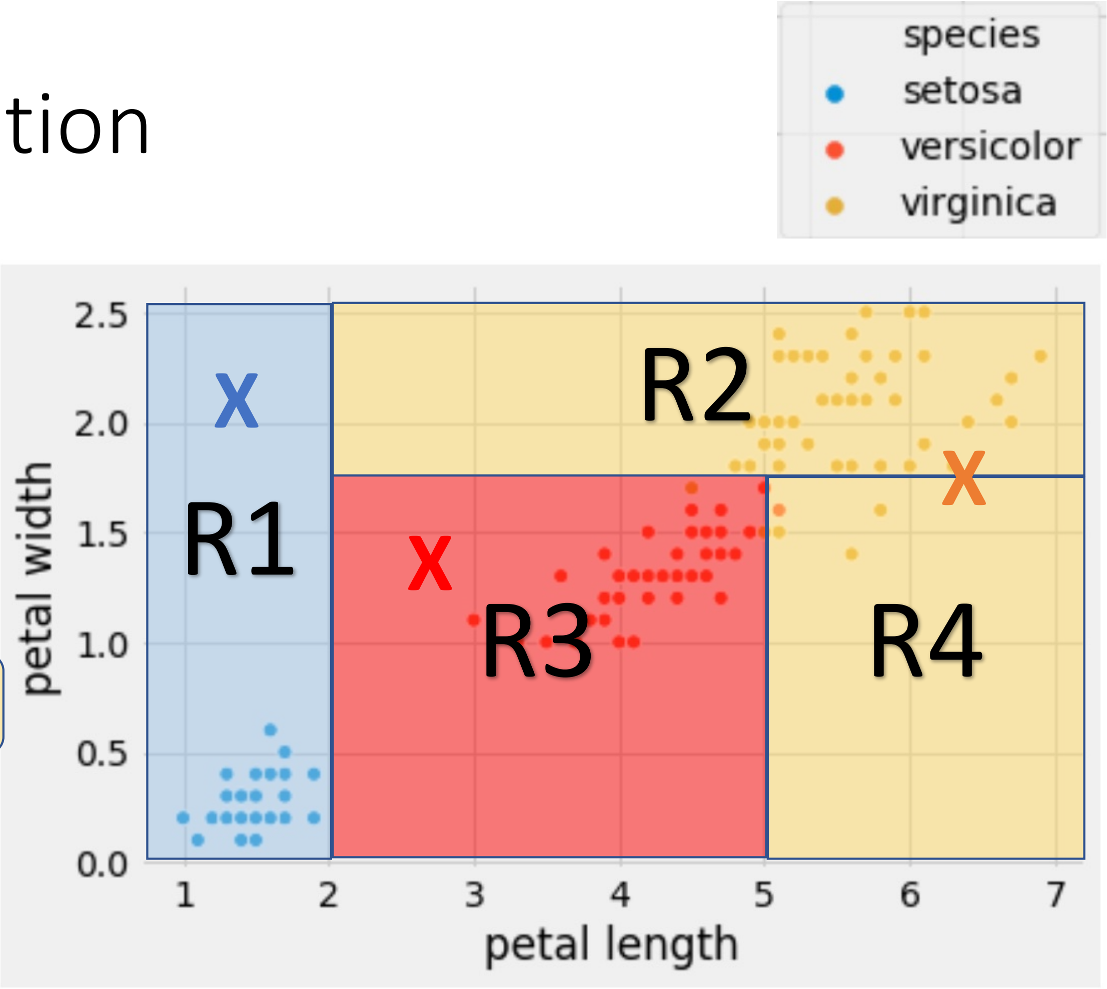
virginica

<=5

>5

versicolor

virginica



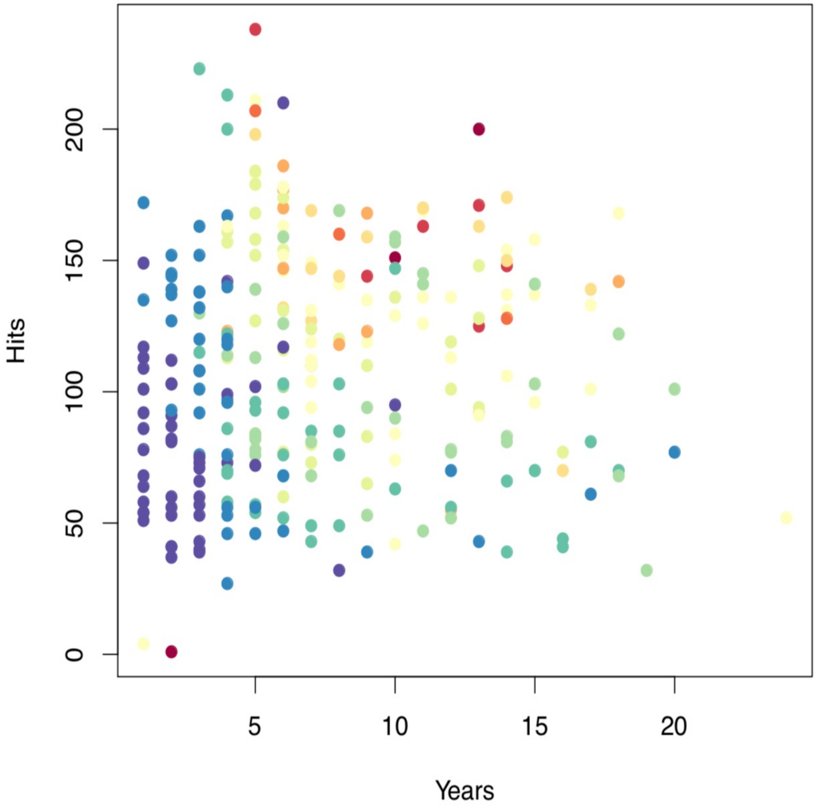
How to grow a tree?

- Partition the feature space into non-overlapping “box” regions
  - predict a response value for each region
- Recursive binary splitting
  - make the optimal split each time by minimizing some loss function

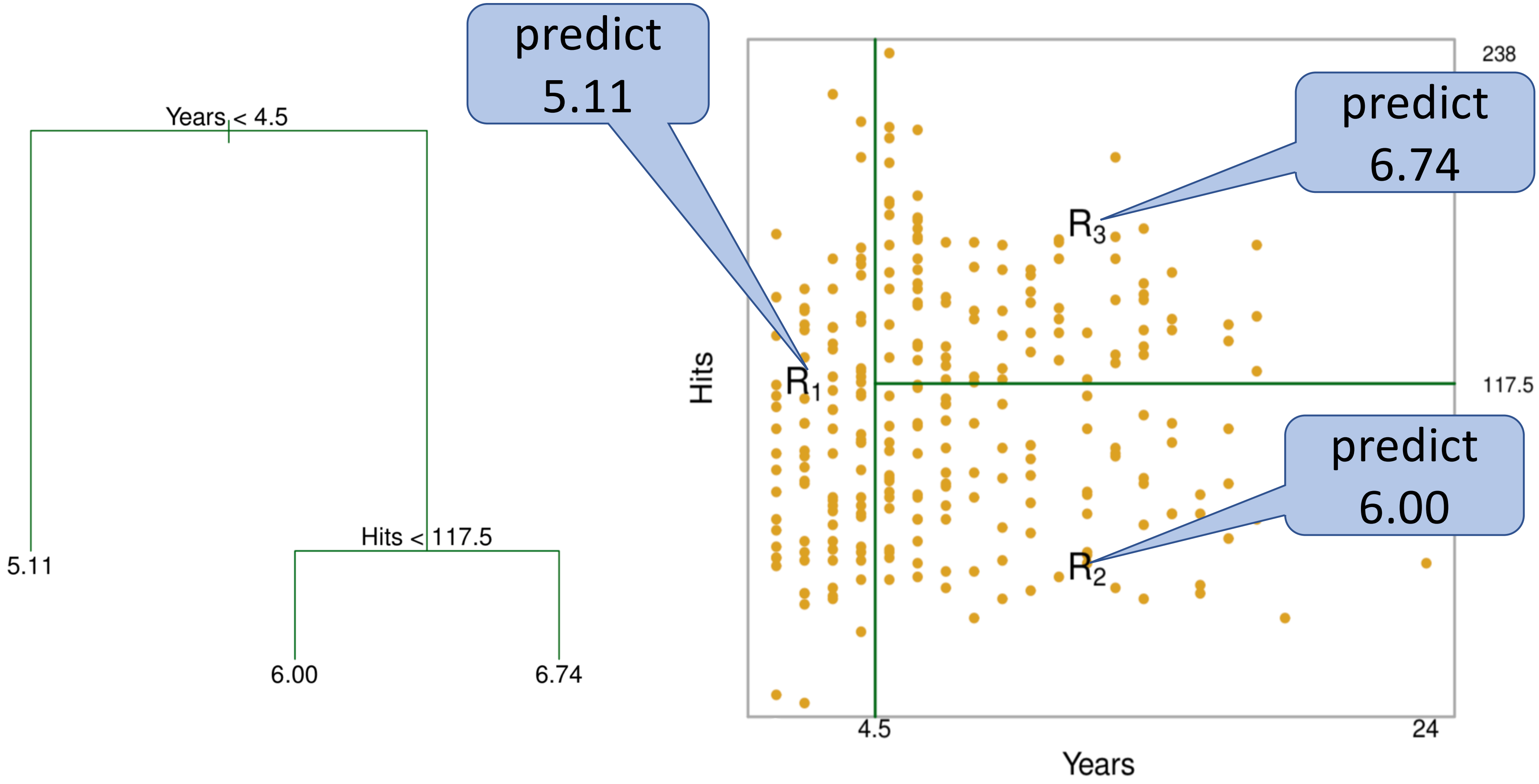
- Any problem?
  - overfitting!!!
  - need a stopping rule: max of splits, min of samples in each leaf region…
  - penalize the size of a tree: grow full and prune back

Example: Hitters data

- Predict salary of baseball players
- With features:
  - number of years in major leagues
  - number of hits made last year
- Salary (log scale) color-coded
  - low: blue, green
  - high: red, yellow



Tree for Regression



predict 5.11

predict 6.74

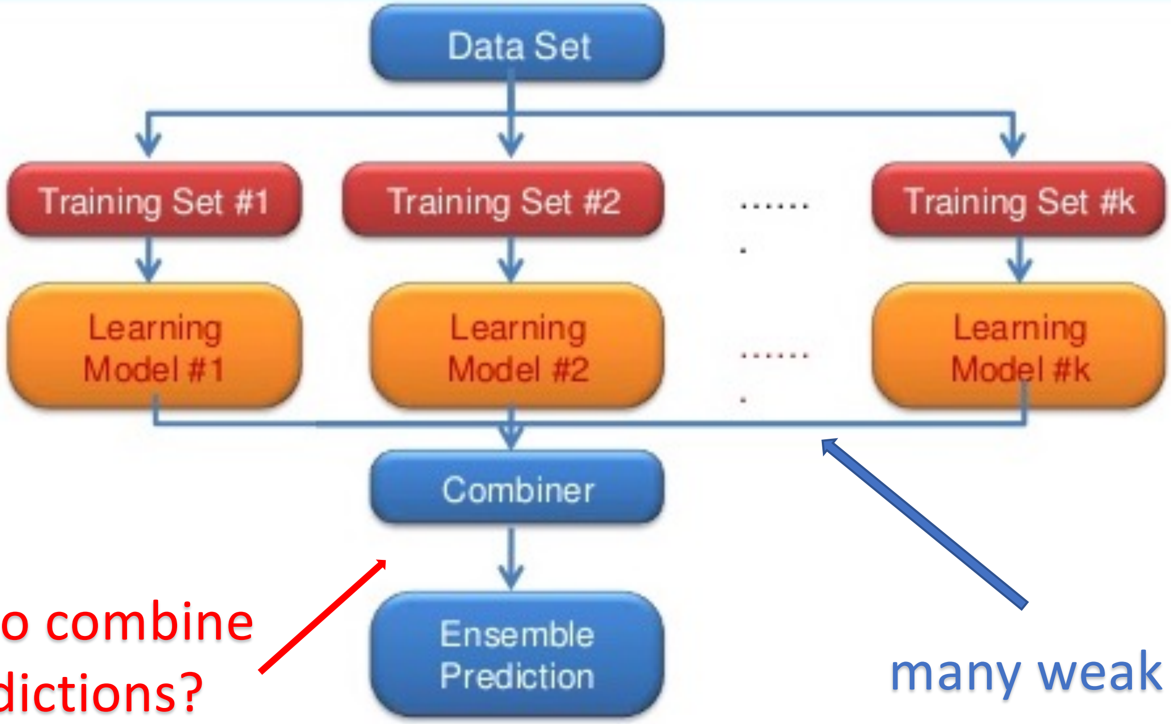
predict 6.00

Pros and Cons

- Easy to display, interpret and explain!
- Very flexible but tend to overfit!
- For both classification and regression!
- Can easily handle:
  - categorical features
  - missing values
- BUT a single tree tend to perform poorly…

Model Ensemble

What is Ensemble?



How to get many training sets?

- Subsampling
- Bootstrap
- Add random noise

many weak learners

How to combine predictions?

- average
- majority votes

one strong learner

Random Forest

- Idea: grow many diverse trees and combine them

- Bootstrap your data:
  - grow one tree for **each bootstrap resample**
  - only allow a **random subset of features** for each split

- Ensemble models: combine predictions from each tree
  - average for regression
  - majority of votes for classification

Takeaway

- Random forest consistently wins
  - robust and require minimum tuning
  - gain prediction accuracy but lose interpretability

- Ensemble revolutionized Machine Learning

- More ensemble examples:
  - bagging
  - boosting
  - model stacking and neural nets (similar idea)