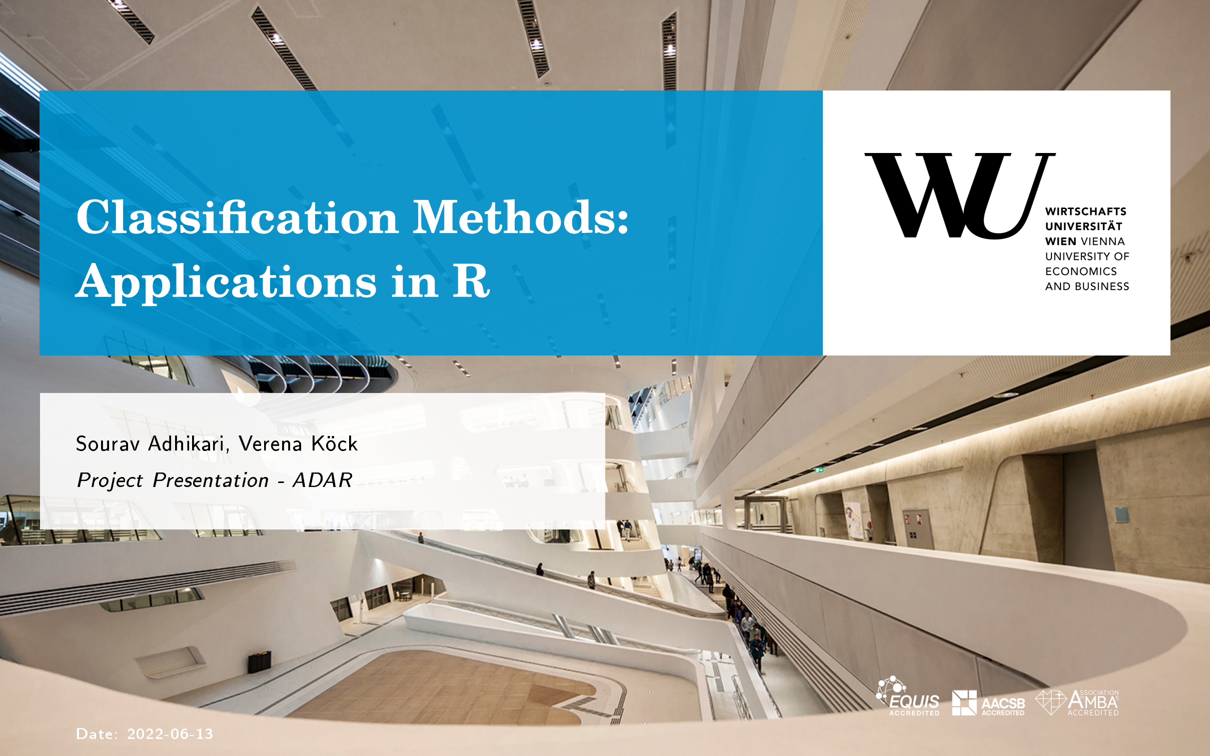


Classification Methods: Applications in R

WU
WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Sourav Adhikari, Verena Köck

Project Presentation - ADAR

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS AMBA ACCREDITED

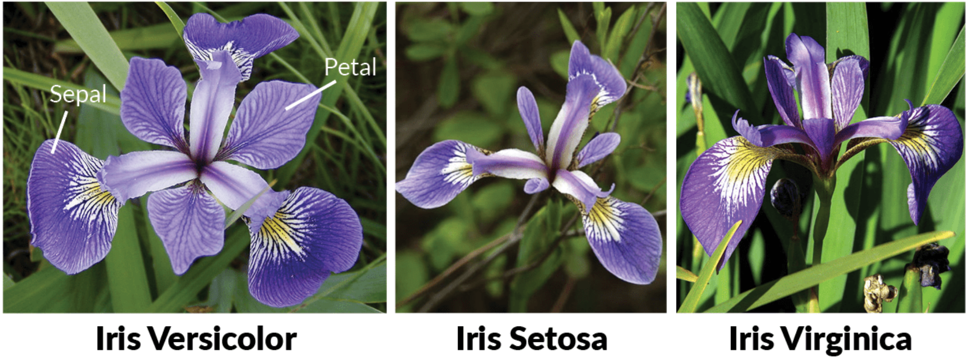
Date: 2022-06-13

(Statistical) Classification: What is it?

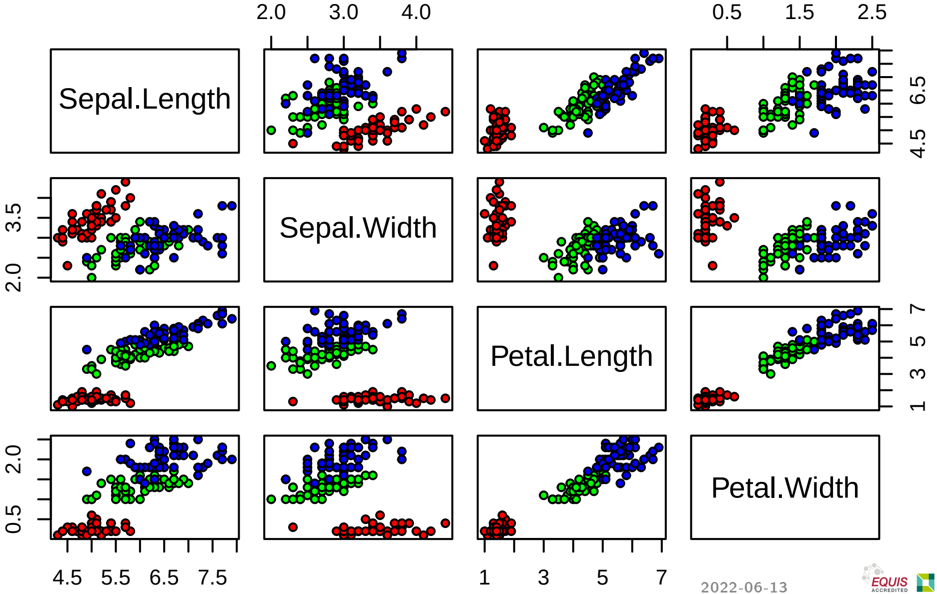
- The problem of identifying which of a set of categories an observation belongs to.
  - E.g. assigning an incoming email to "spam" or "inbox" mailbox.
- Classification can be thought of as two separate problems:
  - binary classification
  - multiclass classification.
- Examples for classification methods are:
  - Naive Bayes
  - k-Nearest Neighbors
  - Neural Networks
  - Others: Decision Trees, Random Forest, Logistic Regression, SVM, etc.
- This project: We explain and present results from first three methods: Naive Bayes, k- Nearest Neighbors and Neural Networks.

The IRIS dataset I

- The data contains 4 measurements for 50 flowers from each of three species of *iris*:
  - `Sepal.Length`, `Sepal.Width`, `Petal.Length` and `Petal.Width` in cm
  - `Species`: setosa, virginica and versicolor



IRIS Data: setosa(red), versicolor(gr.), virginica(bl.)



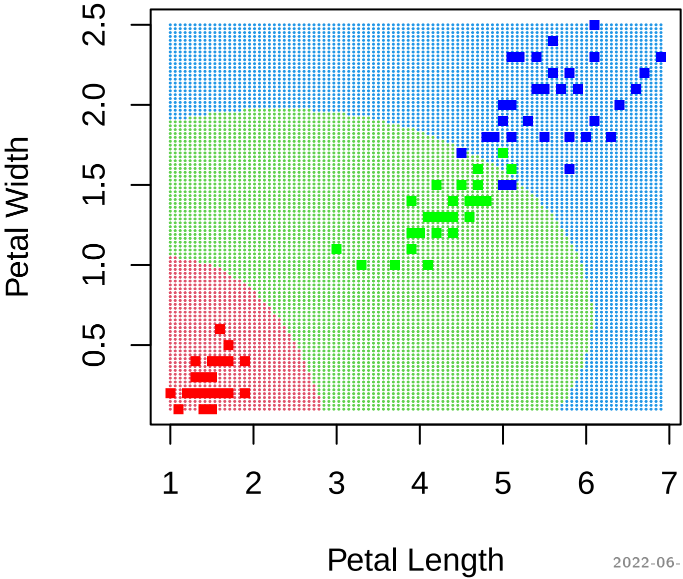
Naive Bayes

- Naive Bayes classifiers are simple "probabilistic classifiers" based on Bayes' theorem.
- *Disadvantage*: (**Strong**) assumption, that the features are independent (i.e presence of one particular feature does not affect the other). Hence the adjective **naive**.
- *Advantage*: Requires only a small number of training data to estimate the parameters.
- Let y be the category variable, and X the features, then Bayes theorem is:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)},$$

- Steps:
  1. Estimate prior probability $P(X)$: Compute the relative frequency of each class/species.
  2. Assume normal distribution for each class (species). Estimate μ and σ^2 for each class.
  3. For a new observation, apply Bayes theorem (and normalize) to get a vector of probabilities, e.g. $(\mathbf{0.5}, \mathbf{0.25}, \mathbf{0.25})$!

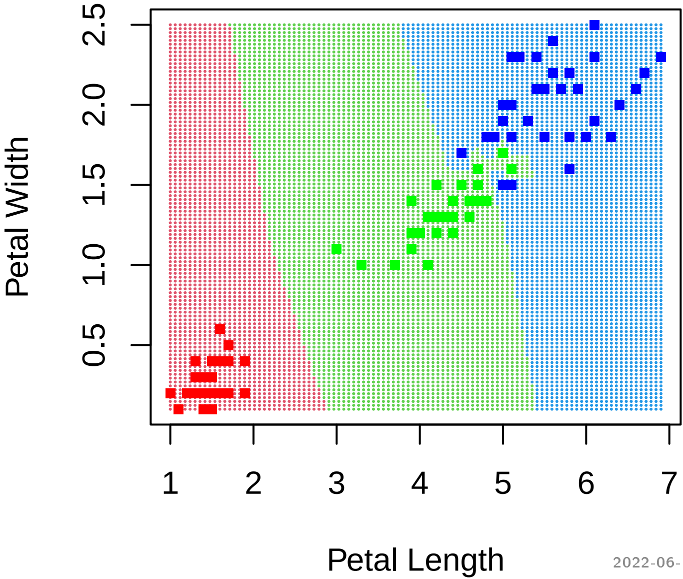
Naive Bayes in R



K-nearest neighbors

- A non-parametric supervised learning method
- Uses a distance metric to make classifications or predictions about the grouping of an individual data point.
- Object is assigned to the class it is most common with among its k nearest neighbors.
- *Advantages*: Easy to understand and implement, no assumptions required
- *Disadvantages*: Curse of Dimensionality

1-nearest neighbour


Petal Width
Petal Length