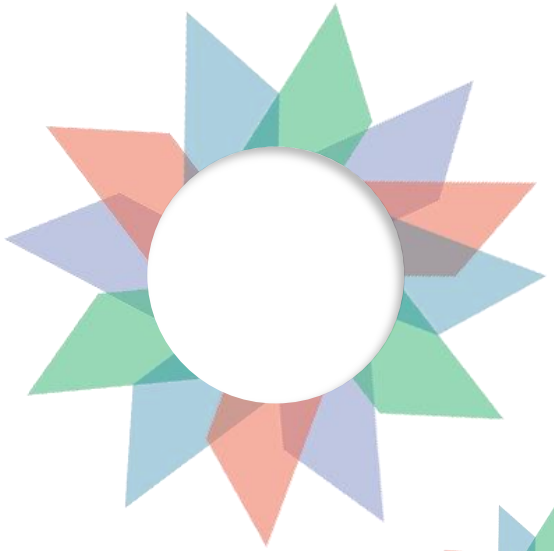
# Bootstrap methods for small sample modeling

**Elena Ballante**
**Raffaella Cabini**
**Chiara Bardelli**

# Table of contents

# Introduction

- Methods to increase stability
- Small sample size problem
- Methods to extract as much information as possible

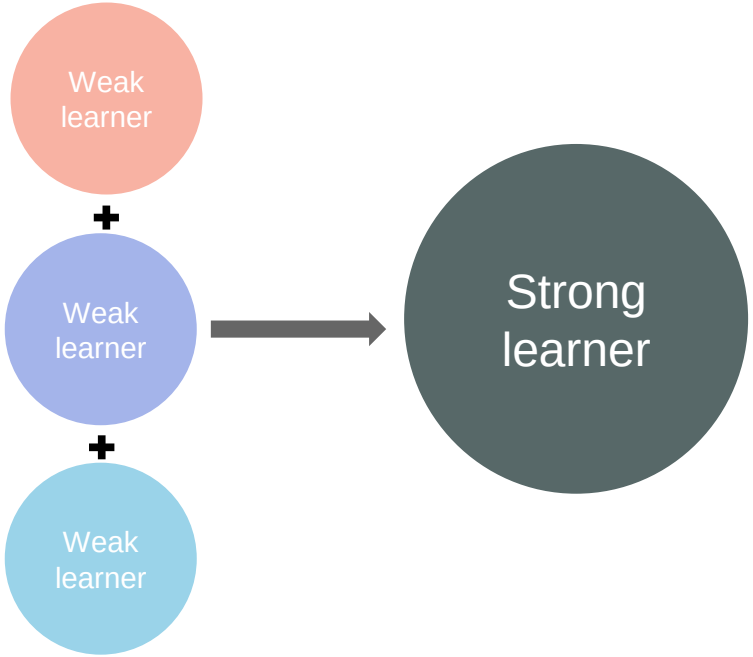


# Ensemble modeling – Bagging

# Ensemble models

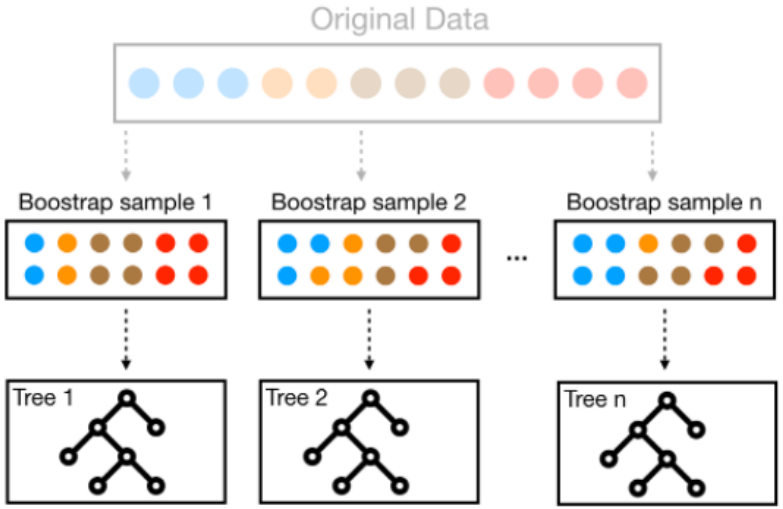
Ensemble models combine weak learners (models) to create a stronger one to reduce the prediction error

- Weak learner
  - Only slightly better than random guess
  - Very easy to find
- Strong learner
  - Very accurate
  - High computational cost



# Bagging (Bootstrap Aggregating)

1. Sample records with replacement (aka "bootstrap" the training data)
2. Fit an overgrown tree to each resampled data set
3. Average predictions

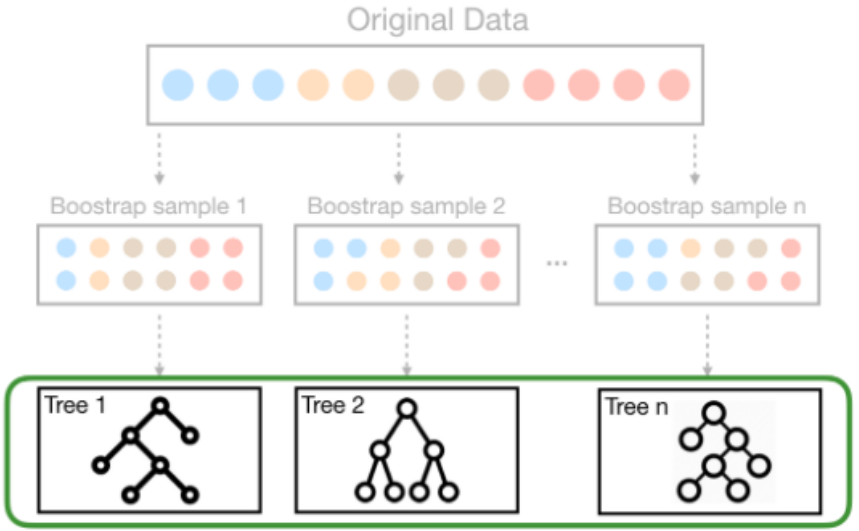



Trees can be highly correlated because they use same variables on similar datasets

# Split-variable randomization

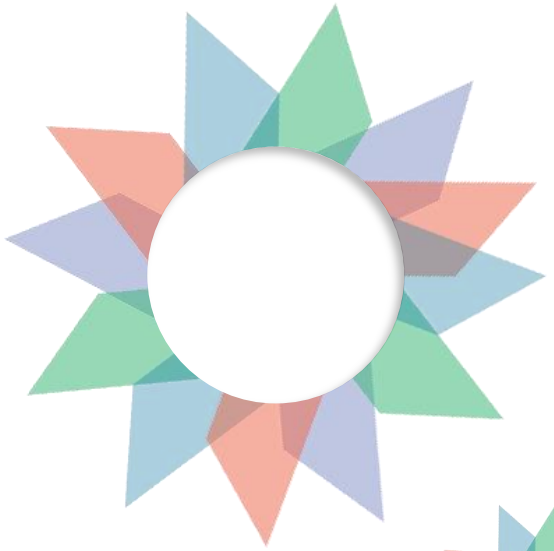
Follow a similar bagging process but each time a split is to be performed, the search for the split variable is limited to a random **subset m** of the **p variables**:

- regression trees: m=p/3
- classification trees: m=sqrt(p)
- m is commonly referred to as ***mtry***



# Advantages

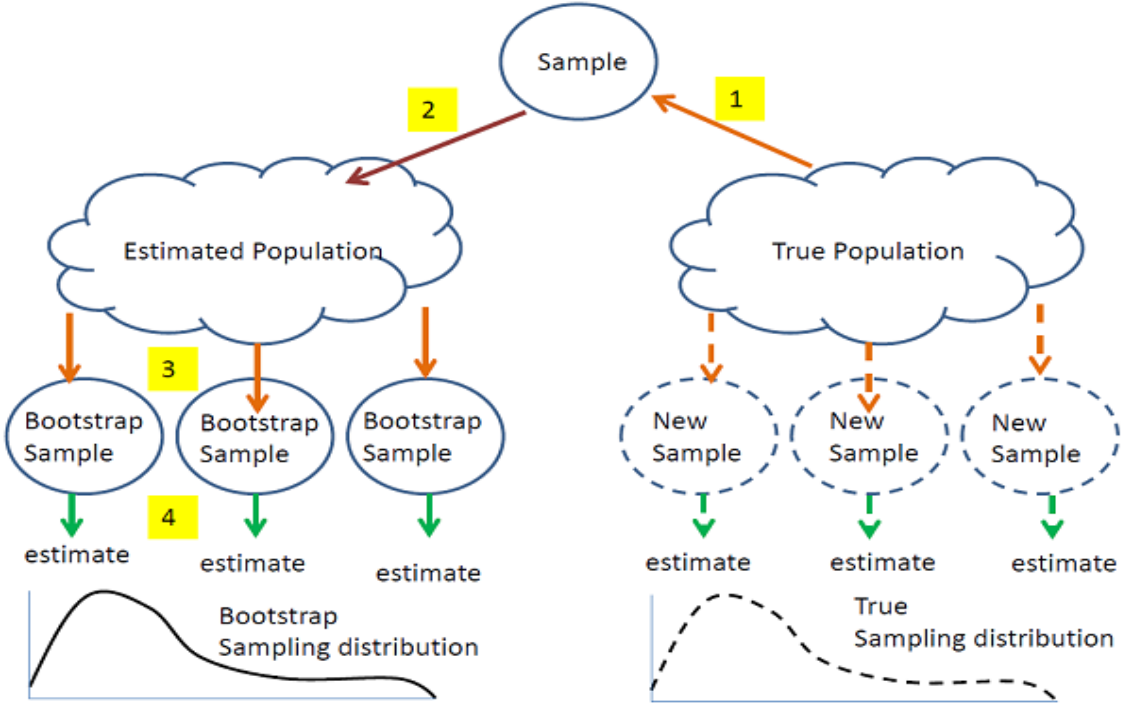
- Avoid the overfitting problem
- It produces a highly accurate classifier and learning is fast
- It can handle high number (thousands) of input variables without variable deletion.
- No assumptions on data distribution



# Bootstrap methods

# Efron’s Bootstrap (1979)

The original bootstrap method assumes that the sample has the same relationship to the popu lation as it has to an empirical distribution that is created by resampling with replacement from the original distribution N samples of the same size as the original sample.



# Efron’s Bootstrap (1979)

Define a random sample of size n drawing with replacement from the original dataset. The new sample is called bootstrap sample.

The bootstrap samples are used to train the ensemble models.

This is equivalent to associate to data points a vector of weights $(\pi_1, \pi_2, \ldots, \pi_n)$ where $\pi_i = c_i/n$ and

$$(c_1, c_2, \ldots, c_n) \sim Multinom(n, (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}))$$

# Rubin’s Bootstrap (1981)

The Rubin's bootstrap, also called Bayesian bootstrap, modifies the Efron's bootstrap defining the vector of weights as

$$(\pi_1, \pi_2, \ldots, \pi_n) \sim Dirichlet(1,1,\ldots,1)$$

In Taddy et al (2015), the authors introduce the idea of replacing Efron's bootstrap with Rubin's bootstrap in bagging algorithm, defining the Empirical Bayesian Forests

**Algorithm:**

1. Draw $(\pi_1, \pi_2, \ldots, \pi_n) \sim Dirichlet(1,1,\ldots,1)$
2. Run a weighted-sample decision tree

Repete B times (B is the number of the final trees)

**Two main drawbacks of Efron’s and Rubin’s bootstrap:**

- No prior opinions are taken into account
- Inference and prediction are based only on observed values

# Proper Bayesian Bootstrap (1996)

Proper Bayesian bootstrap was proposed in Muliere and Secchi (1996). The basic idea is to sample both from the empirical and from the prior distribution.

Suppose we have $X_1, \ldots, X_n \; i.i.d. \sim F$ $\quad F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq x]$

**Aim:** given a random variable $R(X, F)$ estimate the distribution of $R$

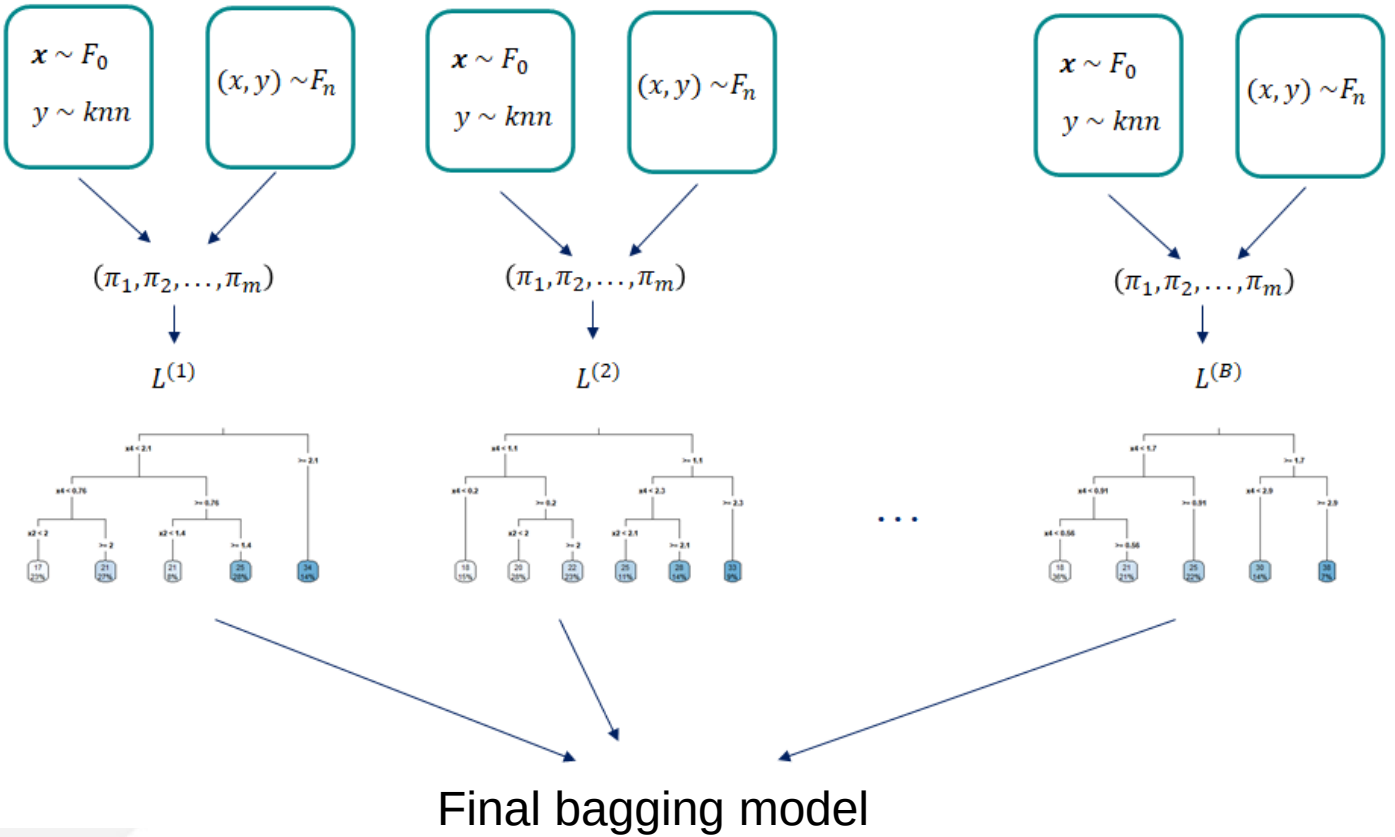
**Procedure:**

1. Generate m observations $x_1^*, x_2^*, \ldots, x_m^*$ from $(n+k)^{-1}(kF_0 + nF_n)$ (bootstrap sample)
2. Generate a vector of weights $(\pi_1, \pi_2, \ldots, \pi_m) \sim Dirichlet(1,1,\ldots,1)$
3. Compute the quantity $R^* = R(\boldsymbol{X}^*, \boldsymbol{\pi}, F_n)$

Repete this procedure B times and evaluate the bootstrap distribution of $R^*$

**Is it possible to include this type of bootstrap in a bagging algorithm?**

# Generalized Bayesian ensemble trees




Galvani et al 2021

# Conclusions

Advantages:

- Ensemble model generates a stronger classifier that reaches in general better performances
- Introducing the proper Bayesian Bootstrap in bagging algorithms seems to stabilize the prediction model (especially in case of low sample size)
- Opinion of experts can be taken into account in generating the bootstrap sample

# Thank you