UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Random Forests

Aruni Yasmin Azizah*, Adila Alfa Krisnadhi, Siti Aminah,
Dina Chahyati, Fariz Darari

CSGE603130: Kecerdasan Artifisial dan Sains Data Dasar
Gasal 2022/2023

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Outline

1. Ensemble Learning
2. Random Forest

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Ensemble Learning

Sumber:

- Adila A. Krisnadhi, Slides Materi Machine Learning “CART & Random Forests”, Semester Genap 2020/2021

FAKULTAS ILMU KOMPUTER
UNIVERSITAS INDONESIA

Ensemble Learning

Prinsip: prediksi ditentukan dengan mengambil rata-rata dari beberapa model yang dibangun (mengurangi variance)

$$f(y|\mathbf{x}) = \frac{1}{|M|} \sum_{m=1}^{M} f_m(y|\mathbf{x})$$

(base) model ke-m

- bias akan menyamai base model tetapi dengan variance lebih rendah
- Untuk classification, gunakan **committee method**, yaitu ambil vote mayoritas sebagai hasil prediksi
- … atau gunakan metode *stacking*

$$f(y|\mathbf{x}) = \sum_{m=1}^{M} w_m f_m(y|\mathbf{x})$$

setiap base model diberikan bobot (weight)

Bagging (Bootstrap Aggregating)

- metode **ensemble** sederhana
- Lakukan fitting M base model berbeda kepada sample data yang berbeda-beda.
  - tiap base model kemungkinan akan melakukan prediksi yang bervariasi.
  - data untuk tiap base model diambil dengan random sampling dengan pengembalian (i.e., *bootstrap sampling*) sampai tiap model di-train dengan N examples (N ukuran training set asli).
- Rata-rata, setiap base model akan melihat 63% dari training set asli.
  - Data tersisa 37% dinamakan sebagai *out-of-bag instances*. Dapat digunakan sebagai testing set untuk estimasi kinerja model.
- Kelebihan bootstrap:
  - prevents the ensemble from relying too much on any individual training example, hence enhancing robustness and generalization.
- Bagging tidak selalu meningkatkan kinerja model

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Random Forests

Sumber:

- A Hands on Tutorial to Data Science, Chirag Shah, A, 2020
- Slides Sains Data, “Random Forest”, Semester Genap 2020/2021
- Kevin P. Murphy, “Probabilistic Machine Learning: An Introduction”, MIT Press, 2021.
- Random Forests Video by StatQuest with Josh Starmer [StatQuest, 2022]

Random Forests

- Model yang dihasilkan oleh algoritma decision tree mudah terjadi overfit pada data.
- Tapi… decision tree adalah model yang mudah diinterpretasi, mudah dibangun dan masih banyak kelebihan lainnya yang dimiliki!
- Why not make many?
  - Bangun beberapa decision tree (a forest!), setiap tree mempunyai struktur yang sedikit berbeda.
  - Untuk setiap tree, bagaimana prediksi yang dilakukan oleh decision trees tersebut?
  - Tetapkan prediksi akhir berdasarkan prediksi-prediksi yang dilakukan oleh setiap tree, misal melalui voting atau ambil mean-nya.
- Salah satu metode **ensemble**.
- Random forest berusaha memperbaiki kecenderungan overfitting yang dilakukan oleh decision tree pada training set

Random Forests

Untuk training set dengan ukuran N, setiap decision tree dibangun berdasarkan langkah-langkah berikut ini:

1. sebuah sampel diambil secara random dengan pengembalian dari N examples pada training set asli. Sampel ini digunakan sebagai training set untuk membangun sebuah decision tree.
2. Bila dataset mempunyai D variabel input, pilih bilangan bulat d (usahakan d jauh lebih kecil dari D) sedemikian sehingga, pada setiap node ketika membangun tree, d variabel input dipilih sebagai kandidat splitting point dari D kemungkinan variabel input pada dataset asli.
3. Lakukan langkah 1-2 sampai semua tree dibangun.

Prediksi random forest adalah dengan mengagregasi prediksi dari semua tree yang dibangun.

Contoh: Random Forest Learning

Step 1: Buatlah Bootstrap Dataset

Original Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

Ambil 1 data secara acak sebanyak N kali (4 kali) dengan pengembalian

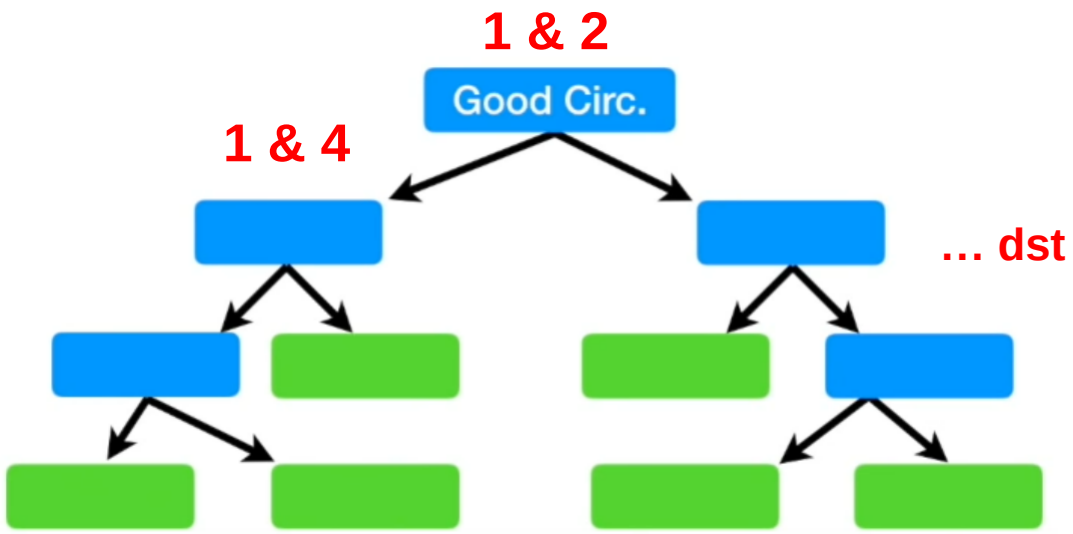
Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Sangat mungkin bila ada data duplikat!

Contoh: Random Forest Learning

Step 2: Bangun Decision Tree **tapi…** untuk menentukan atribut pada splitting point, ambil d atribut saja (secara acak → random subset!) sebagai calon splitting point.



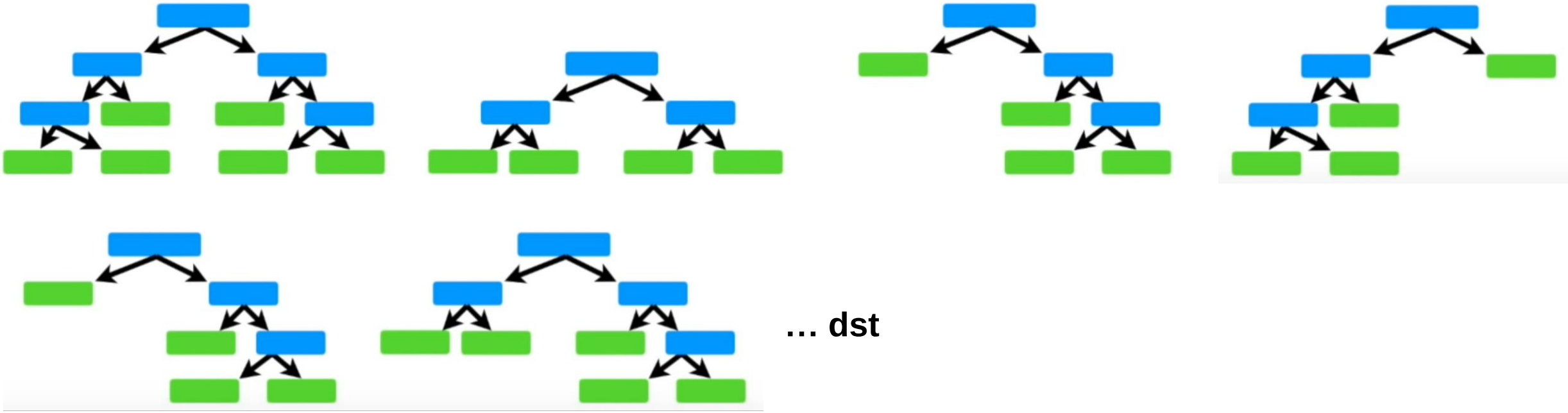
Bootstrapped Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Attributes: 1 2 3 4

Contoh: Random Forest Learning

Step 3: Balik ke Langkah 1 dan 2 untuk membentuk bootstrap dataset dan membangun decision tree lagi. Terus bangun decision “tree” sampai jumlah yang sudah ditentukan sehingga membentuk sebuah “forest”.



… dst

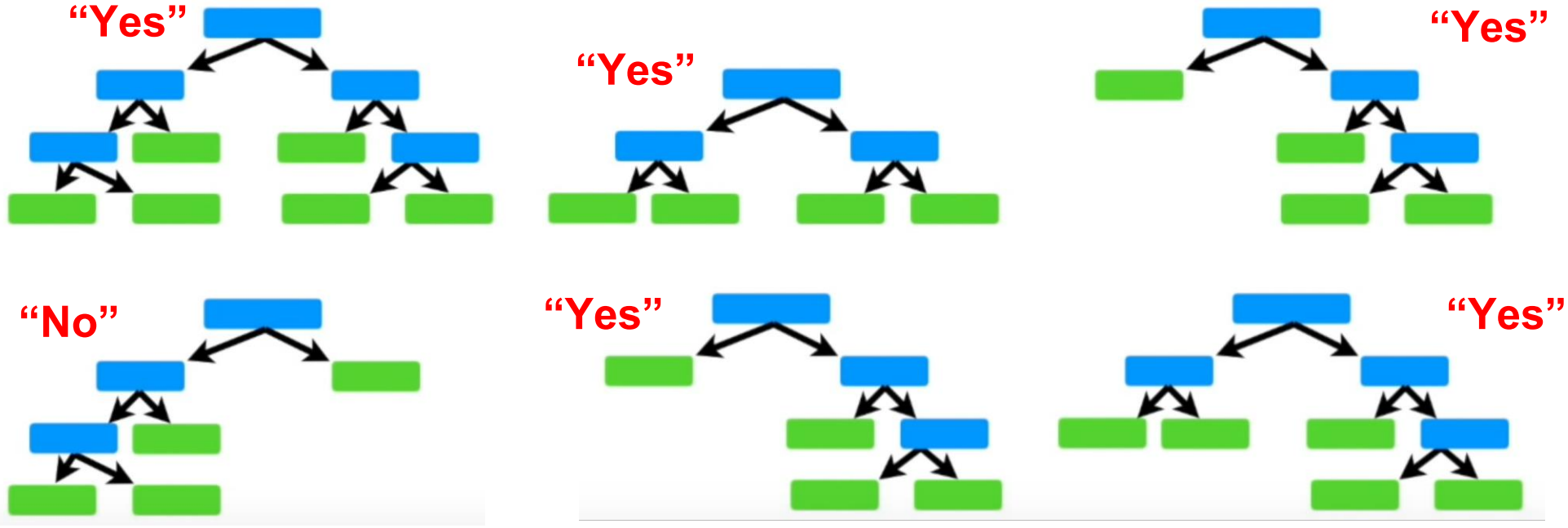
*agar kasusnya sederhana, asumsikan dalam contoh ini, forest terdiri dari 6 tree saja :D

Contoh: Random Forest Testing

- Untuk sebuah data, setiap tree harus melakukan prediksi nilai output dari data tersebut

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	"Yes"

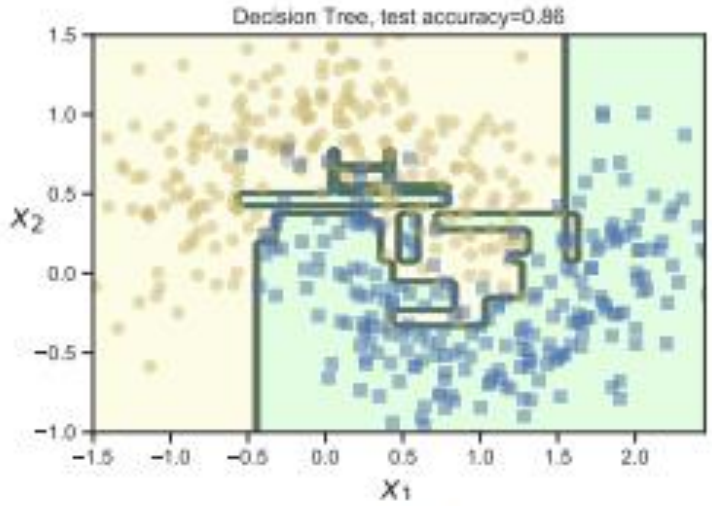
Heart Disease	
Yes	No
5	1



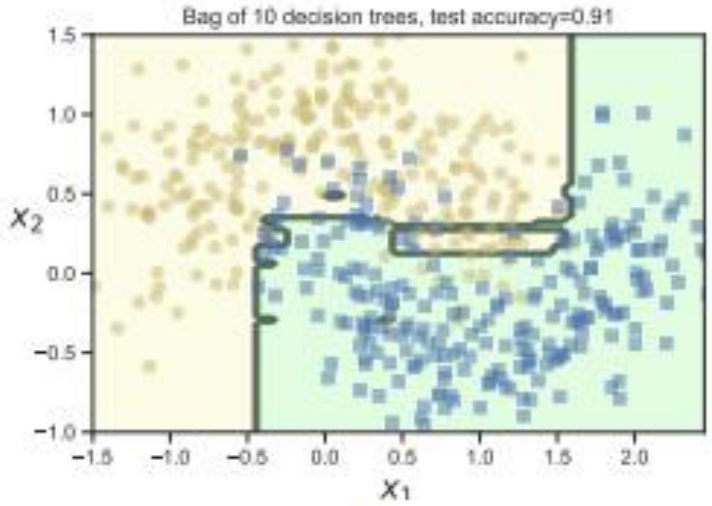
Random Forest Hyperparameters

- Testing dapat digunakan untuk menentukan bagus atau tidaknya Random Forest yang dibangun (*ideally cross-validation, but we will explain that later…*)
- Beberapa hyperparameter yang dapat di-tuning
  - n_estimators: banyaknya tree dalam forest
  - max_samples: ukuran bootstrap sample yang dibuat untuk training setiap tree
  - max_depth
  - min_samples_split
  - min_samples_leaf
  - max_features
  - max_leaf_nodes
  - dll.

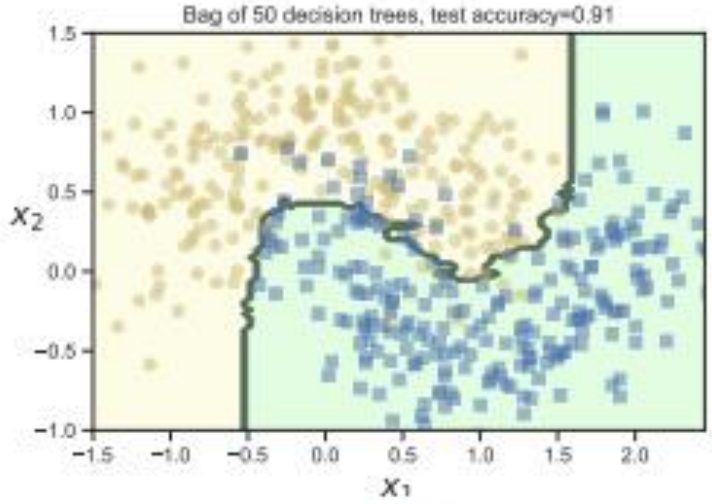
Random Forest



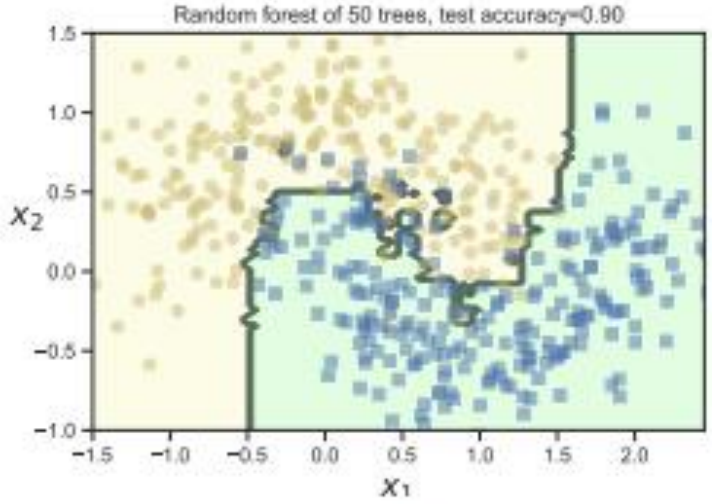
(a)



(b)

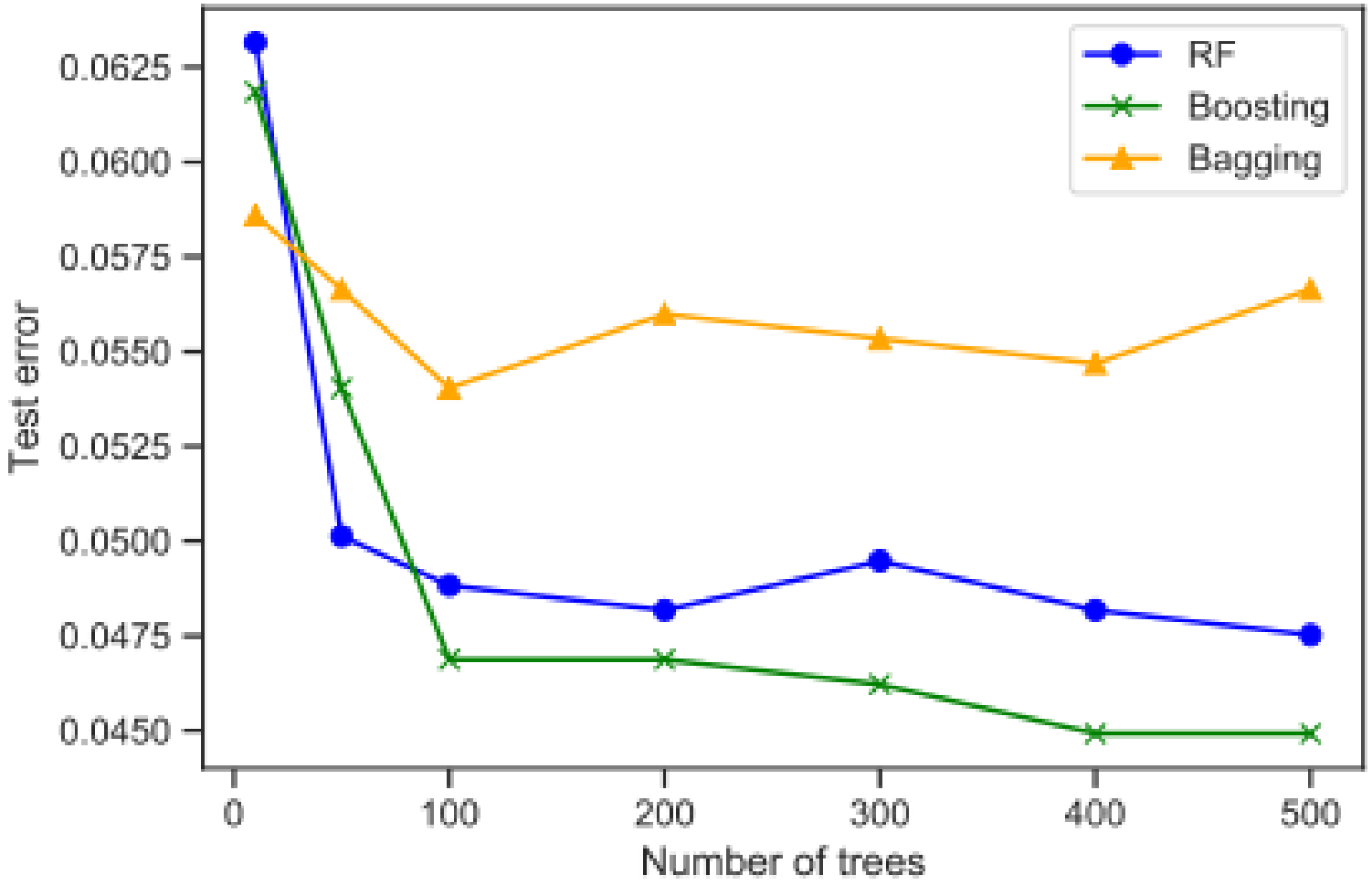


(c)



(d)

Random Forest



Random Forest

- Jadi..., mengapa random forest dapat bekerja lebih baik dibandingkan dengan sebuah decision tree?
- Walaupun tidak ada kesepakatan yang jelas antara periset, ada dua hal yang diyakini dibalik proses random forest:
  - Seperti kata pepatah, "*Nobody knows everything, but everybody knows something*."
  - Ketika dihadapi dengan "forest of trees", setiap tree bukan model prediksi yang sempurna atau yang paling akurat. Sebagian besar tree memberikan prediksi yang tepat untuk sebagian besar data.

Jadi, walaupun ada tree yang memberikan prediksi yang kurang tepat, mayoritas dari tree dalam hutan akan memprediksi label kelas yang tepat.

Random Forest

- dan karena kita menggunakan mode (vote mayoritas) untuk menentukan kelas pada classification, hasil akhir tidak akan dipengaruhi oleh prediksi-prediksi yang salah. Secara intuitif, kebenaran dari pernyataan ini bergantung pada unsur randomness pada metode sampling.
- Semakin random sampelnya, semakin menurun korelasi antara tree dan semakin kecil kemungkinan suatu tree akan dipengaruhi oleh prediksi yang salah oleh tree lainnya.
- Selain itu, kesalahan yang dilakukan oleh suatu tree mungkin tidak sama dengan kesalahan yang dilakukan oleh tree lainnya (tempat terjadinya error mungkin tidak sama).
- Sekali lagi secara intuitif, pernyataan-pernyataan di atas bergantung pada seberapa random variabel input (atribut) dipilih. Semakin random, semakin kecil kemungkinan tree yang berbeda melakukan kesalahan/error pada tempat yang sama.

Random Forest: Pros and Cons

(+) Versatile uses
(+) Easy-to-understand hyperparameters
(+) Classifier doesn't overfit with enough trees

(-) Increased accuracy requires more trees
(-) More trees slow down model
(-) Can't describe relationships within data

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

TERIMA KASIH

Disclaimer: Figures and content can be originated from other sources on the Web. The purpose of this slide set is educational only.