Modele de tip “Ensemble” partea I
Modelarea de tip bagging. Modelul random forests

ADRIAN OȚOIU

Modele de tip “Ensemble” partea I

Concepte de bază

- Aprecierea performanțelor modelelor. Compromisul bias/variance
- Weak learners, strong learners
- Modelarea de tip ensemble. Concept, elemente de bază, avantaje

Modalitățile de modelare de tip ensemble

Modelarea de tip bagging

- Descriere generală
- Tehnica boostrapping. Elemente

Modele din clasa random forests

- Descriere generală algoritm
- Probleme specifice legate de utilizarea practică

Aprecierea performanțelor modelelor. Compensarea bias/variance

- Modelele de data mining/machine learning cu învățare supervizată au 2 proprietăți esențiale, **bias** și **variance**
- Teoretic, performanța unui model poate fi formalizată din perspectiva acestor componente ale erorii:

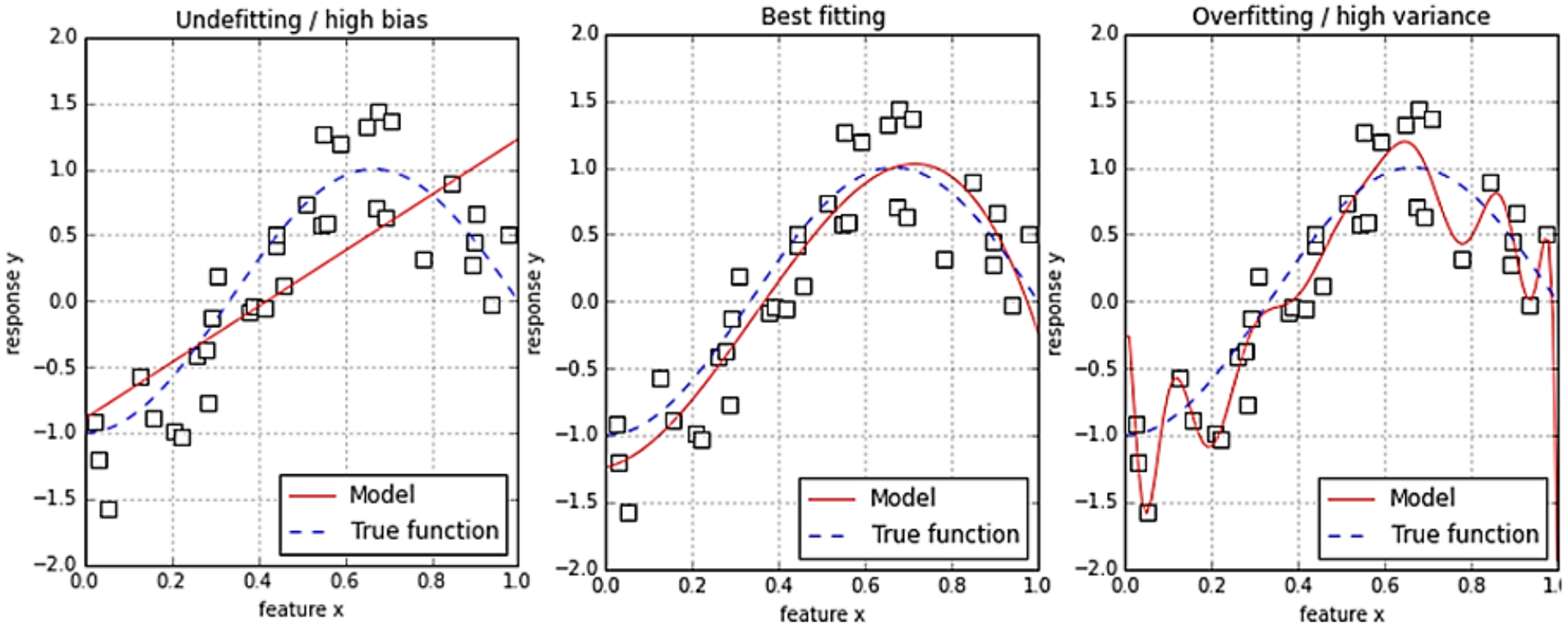
- $E[MSE] = \sigma^2 + (Bias)^2 + Variance$

- unde: -MSE (mean squared error) este suma pătratelor reziduurilor
- σ^2 componenta de 'zgomot alb'; teoretic nu poate fi eliminată
- Bias – arată cât de bine se apropie modelul de relația dintre variabilele explicative (predictori) și variabila dependentă
- Variance- variabilitatea predicției variabilei dependente

Aprecierea performanțelor modelelor. Compromisul bias/variance

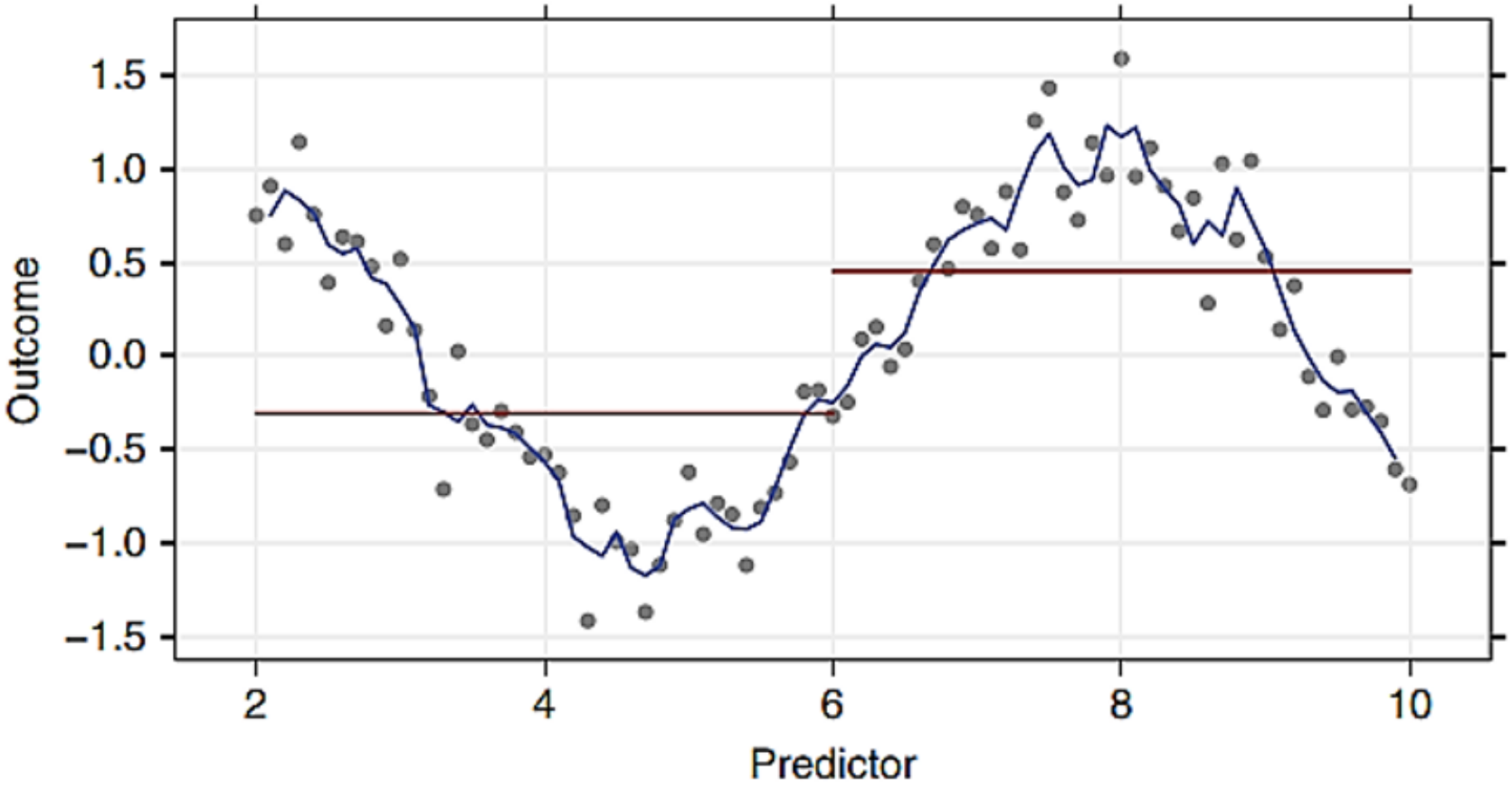
- Importanța celor 2 componente
- **- valori mari pentru bias => underfitting/sub-antrenare:** modelul poate omite relații importante între variabilele predictor și variabila țintă (dependentă).
- - **valori mai pentru varianță** => **overfitting/supra-antrenare:** valorile estimate ale variabilei țintă (modelate) sunt sensibile la variații mici ale valorilor variabilelor predictor, ceea ce poate afecta puterea de predicție a modelelor.
- **Compensarea bias/variance (trade-off):** creșterea valorii unei componente duce la scăderea celeilalte
- Soluția optimă pentru modelare: un model în care ambele componente au valori relativ mici.

Compensarea bias/variance
Exemplul 1



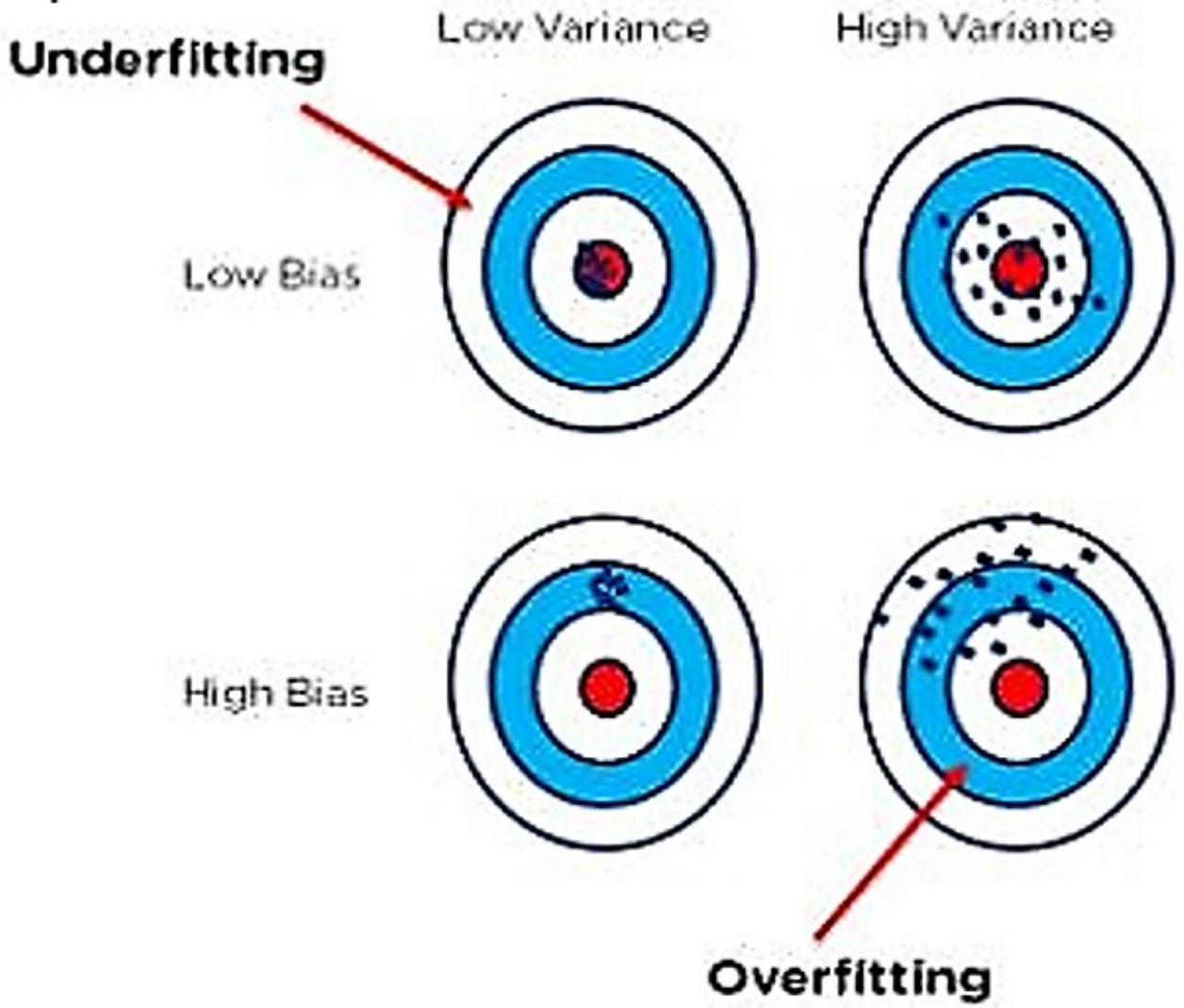
Mueller și Massaron (2016)

Compensarea bias/variance
Exemplul 2



Linia maro- predicție cu bias ridicat si variantă scăzută
Linia albastră- predicție cu variantă scăzută și bias ridicat
Sursa: Kuhn și Johnson (2013)

Compensarea bias/variance
Exemplul 3



Weak learners, strong learners și modelarea de tip ensemble

Weak learners: modele ale căror performanțe sunt slabe. Rezultatele sunt ceva mai bune decât o predicție aleatoare.

Pe acest tip de modele se bazează modelele de tip ensemble, fiind folosite pentru obținerea unui model superior ca performanțe.

Strong learners: modele cu performanțe superioare, ale căror putere de predicție este mult superioară față de cea a unui model aleatoriu.

Printre aceste modele se numără și cele de tip ensemble.

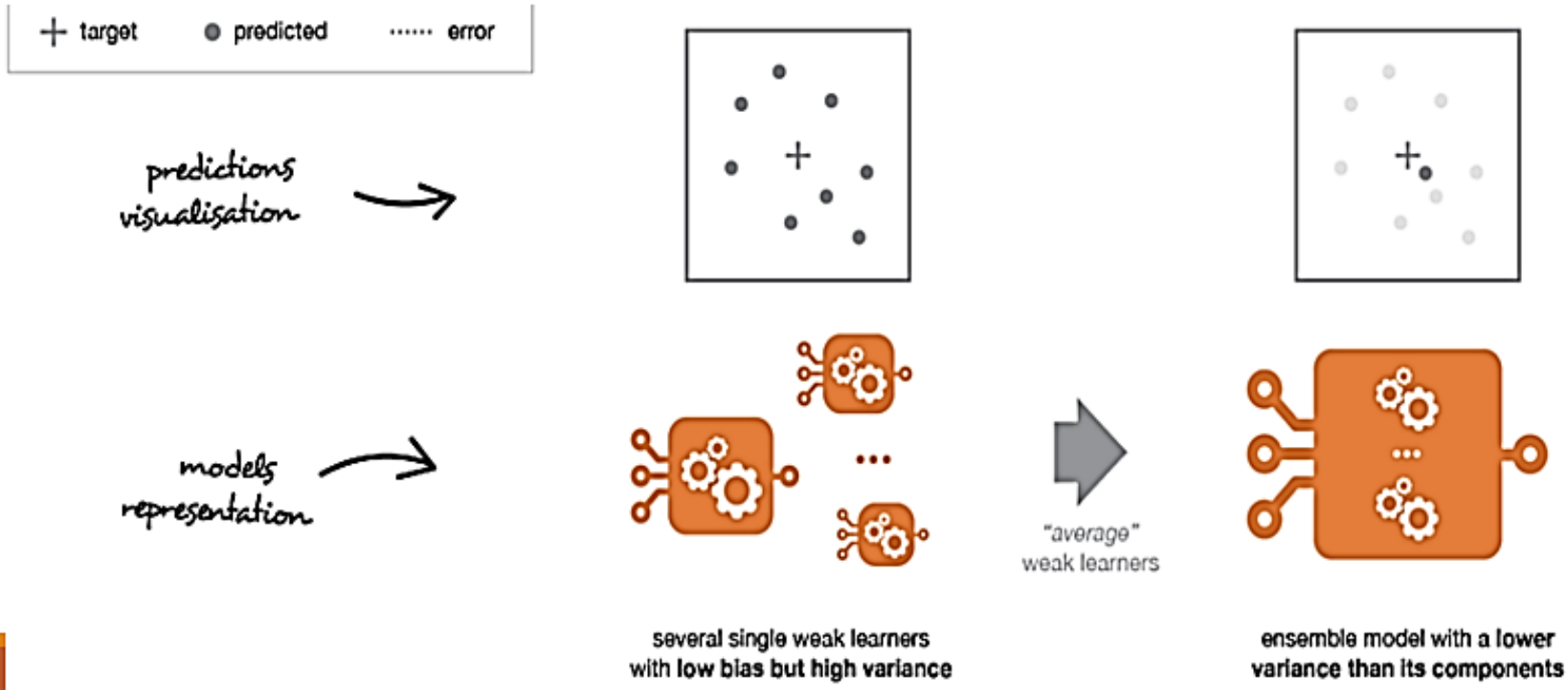
Modelarea de tip ensemble (I)

Obținerea unor modele optime dpdv. al compromisului bias-variance prin modelarea de tip **ensemble**.

Modelarea de tip ensemble (paradigmă de machine learning): tehnica prin care mai multe modele de bază cu performanță scăzută (weak learners), sunt antrenate pentru a rezolva **aceeași problemă**, iar rezultatele acestora sunt combinate pentru a obține modele cu performanțe superioare **(strong learners).**

Exemplu: model ensemble de tip bagging

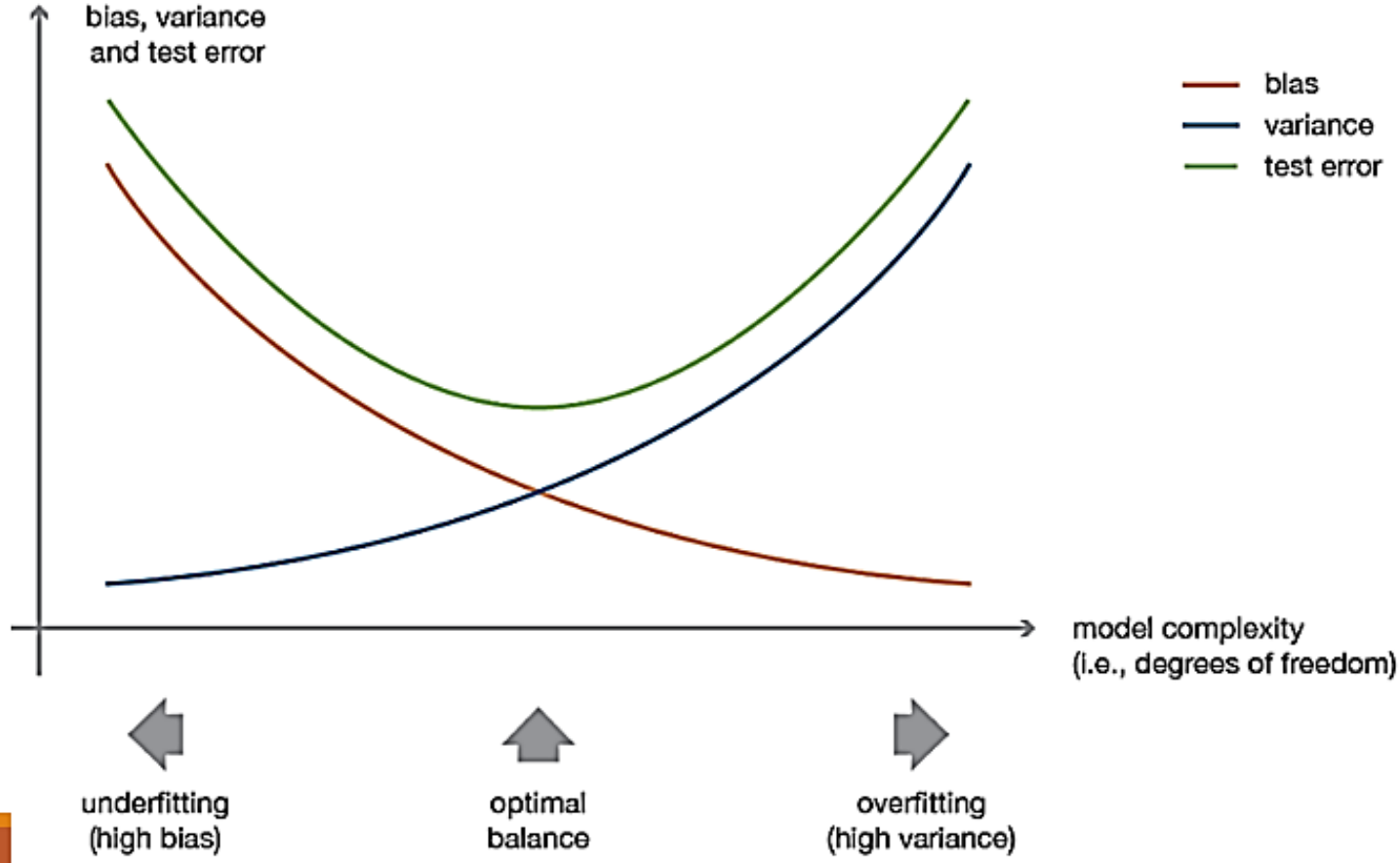
Sursa: Joseph Rocca (2019)



Modelarea de tip ensemble (II)

De regulă se folosește un singur tip de model de bază (de exemplu arbori de decizie), cel puțin pentru algoritmii consacrați (random forests, gradient boosting).

Scopul este estimarea unor modele cu performanță sporită prin obținerea unui compromis optim între bias și varianță, ținând cont de complexitatea modelelor.

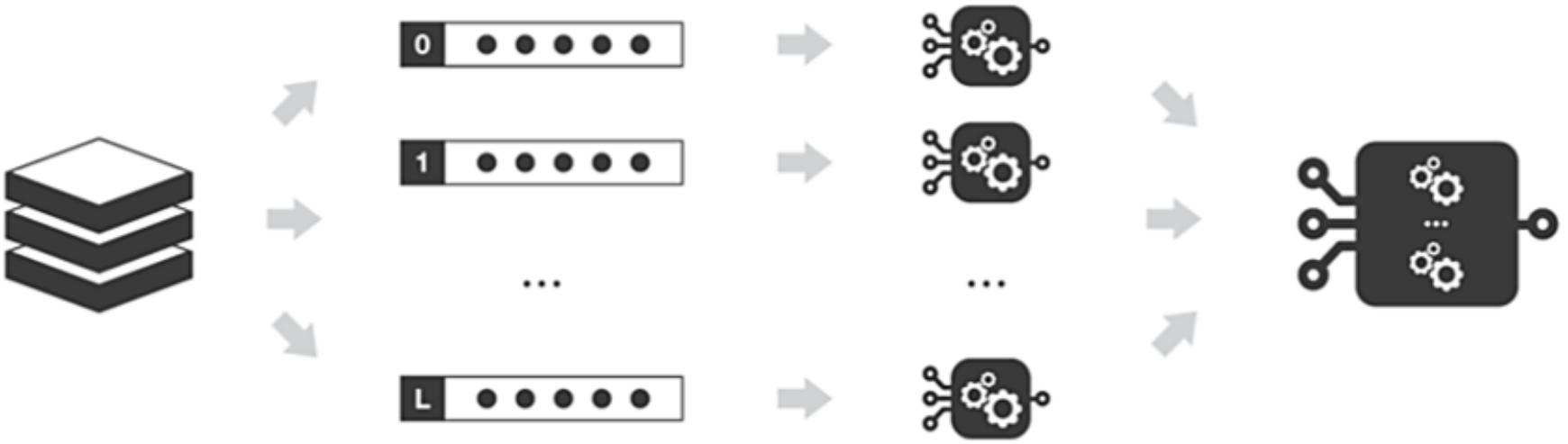


Rocca (2019)

Modele de tip ensemble (III)

Modelarea de tip ensemble s-a structurat în trei mari tipuri/clase de modele

Modele de tip bagging (boostrap **agg**regat**ing)**: folosesc **un singur tip de model de bază** (cel mai adesea arbori de decizie). Algoritmii sunt implementați folosind **estimări independente**, realizate pe eșantioane diferite, iar rezultatele acestora sunt **combinate într-o manieră deterministă** pentru a obține estimarea finală.



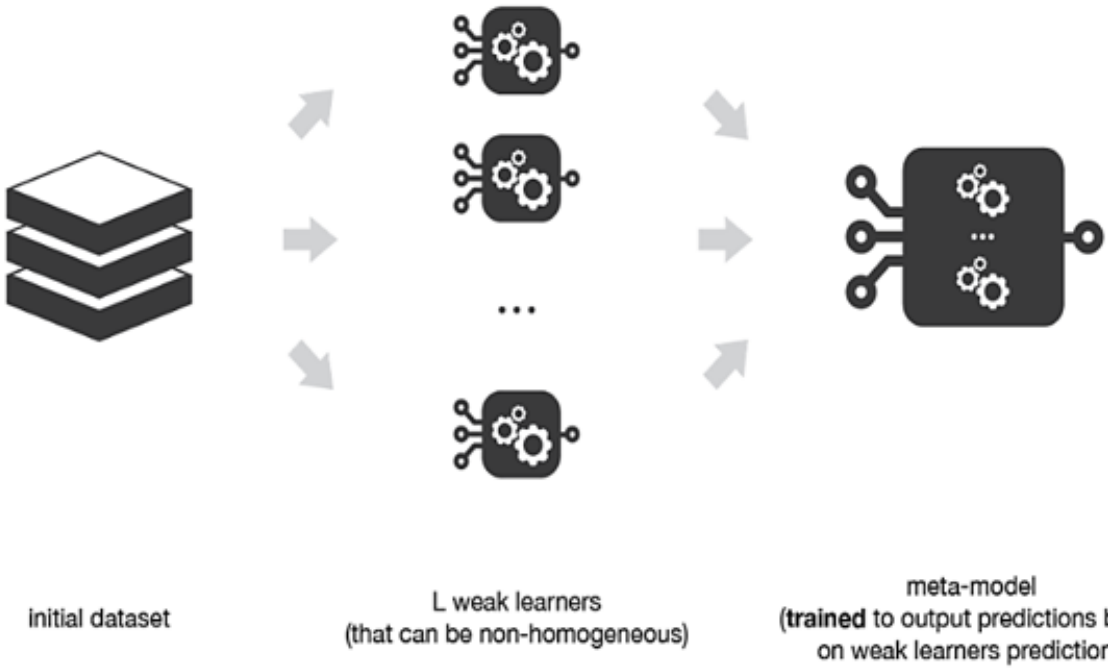
initial dataset | L bootstrap samples | weak learners fitted on each bootstrap sample | ensemble model (kind of average of the weak learners)

Rocca (2019)

Modele de tip ensemble (IV)

Modele de tip boosting: folosesc **un singur tip de model de bază.** Algoritmii sunt implementați într-o **manieră secvențială, adaptivă,** astfel încât estimările generate de către un model de bază depind de estimările modelelor de bază anterioare. Rezultatul final al estimării se obține prin combinarea rezultatelor folosind o metodă deterministă.

Modelele de tip stacking: folosesc de regulă mai multe tipuri de **modele de bază (weak learners).** Aceste modele sunt estimate în paralel, iar pe baza acestora se construiește un meta-model pentru generarea unui rezultat bazat pe predicțiile diferitelor modele de bază.

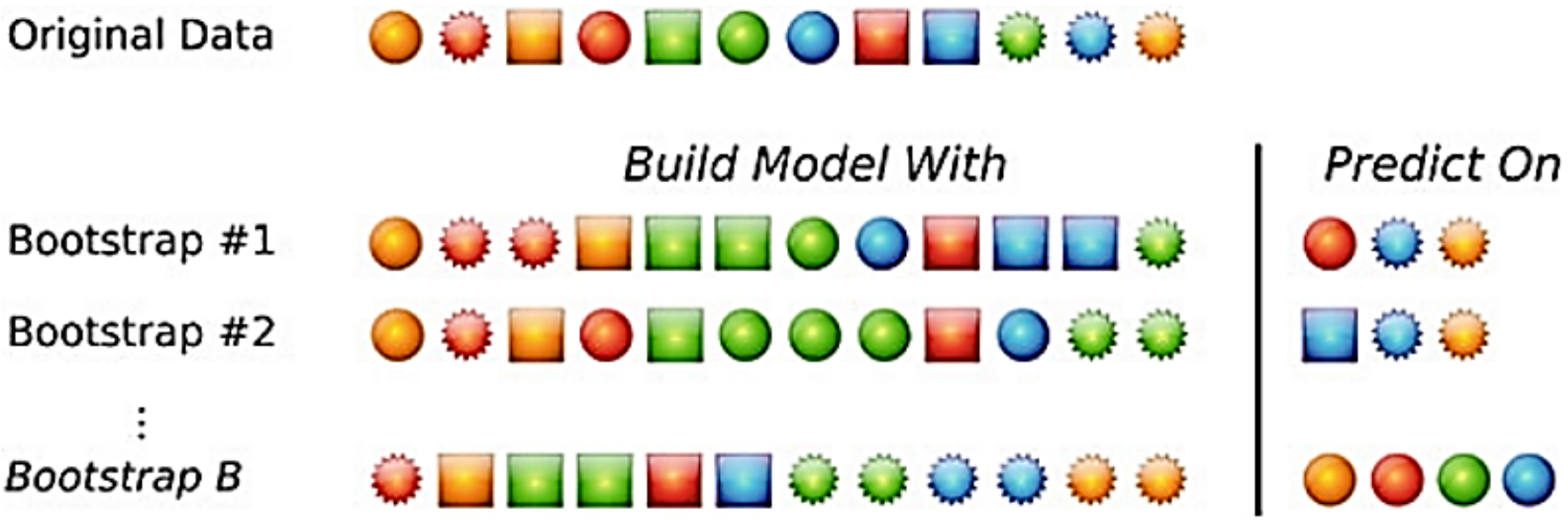


Rocca (2019)

Modele de tip bagging

Cel mai frecvent folosit model de bază sunt arborii de decizie, care se antrenează pe eșantioane distincte, obținute prin metoda **bootstrapping**.

Din eșantionul de date de antrenare de dimensiune N (generat din setul de date inițial) se generează mai multe eșantioane, de dimensiuni egale cu cele ale setului de date inițial.



Kuhn și Johnson (2013)

Eșantionarea cu metoda bootstrap

Precizări legate de metoda bootstrap

- Din setul de date inițial se extrage un eșantion, de regulă prin procedeul cu revenire => o observație se poate regăsi de mai multe ori, chiar în același eșantion.
- Unele observații din setul de date inițial nu se vor regăsi deloc în eșantioanele bootstrap.
- Aceste observații vor fi folosite ulterior pentru validare de tip „out of bag”, utile pentru evaluarea modelului obținut.
- Estimările independente al modelelor de bază se pretează la paralelizare (estimări efectuate simultan pe mai multe resurse informatice).

Modalități de agregare a estimărilor

Rezultatele obținute prin estimarea modelelor de bază se agregă apoi pentru obținerea rezultatelor modelului de tip bagging.

Considerând că s-au estimat L modele de bază, fiecare pe subeșantioane $\bar{x}$ extrase din setul de date de antrenare, iar rezultatele acestora sunt de forma $d_i(x)$, $i \in \{1,L\}$

Rezultatele se agregă în felul următor:

- pe baza mediei in cazul arborilor de regresie (variabilă dependentă de tip continuu)

$$\hat{y} = \frac{1}{L}\sum_{i=1}^{L} d_i(\bar{x})$$

- pentru modelele cu variabilă dependentă de tip discret (cazul arborilor de clasificare) prin vot majoritar.

$$\hat{y} = argmax_{d_i(\bar{x})}\left(d_i(\bar{x})\right)$$

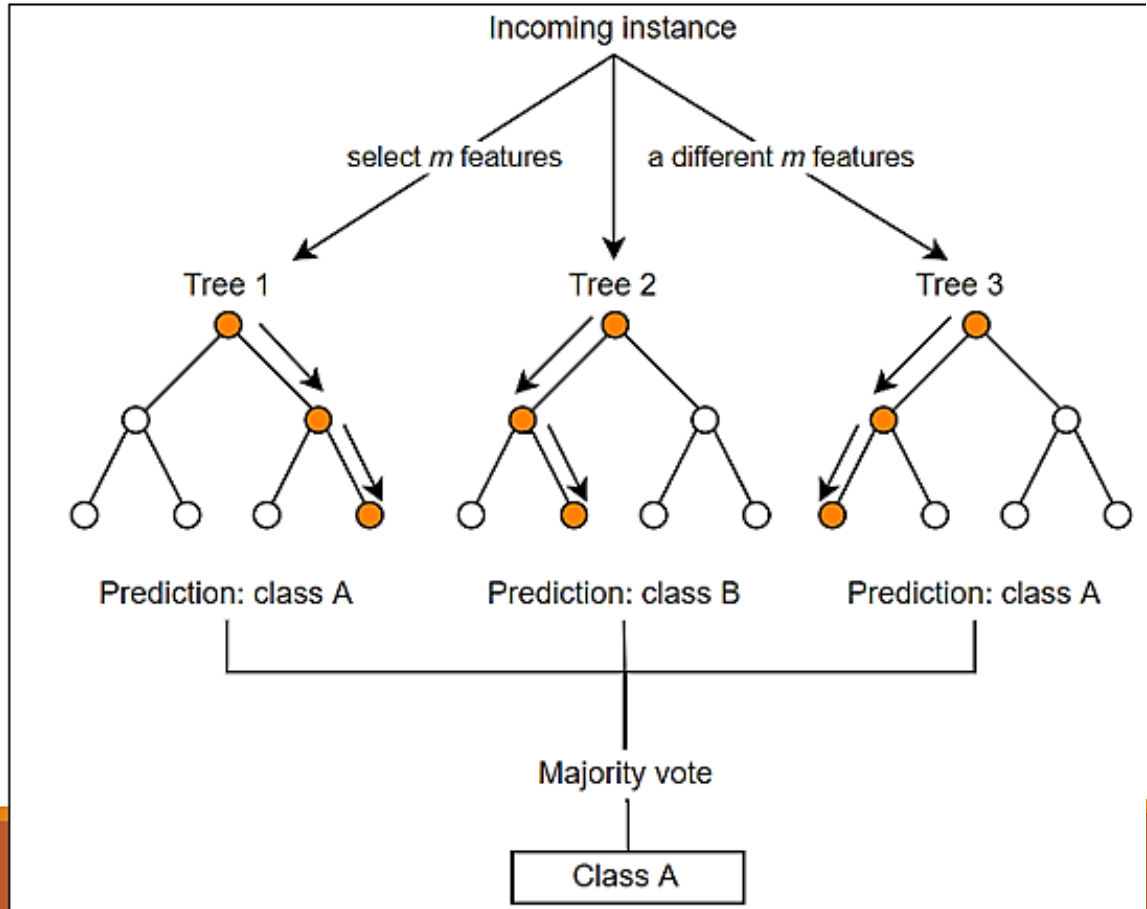
Bonaccorso (2018)

Exemple de agregare a estimărilor

Agregare pe baza mediei

$$\frac{1}{N_{\text{tree}}}\left[\;\;+\;\;+\;\;+\;\;\ldots\;\;\right]$$

Agregare prin vot majoritar

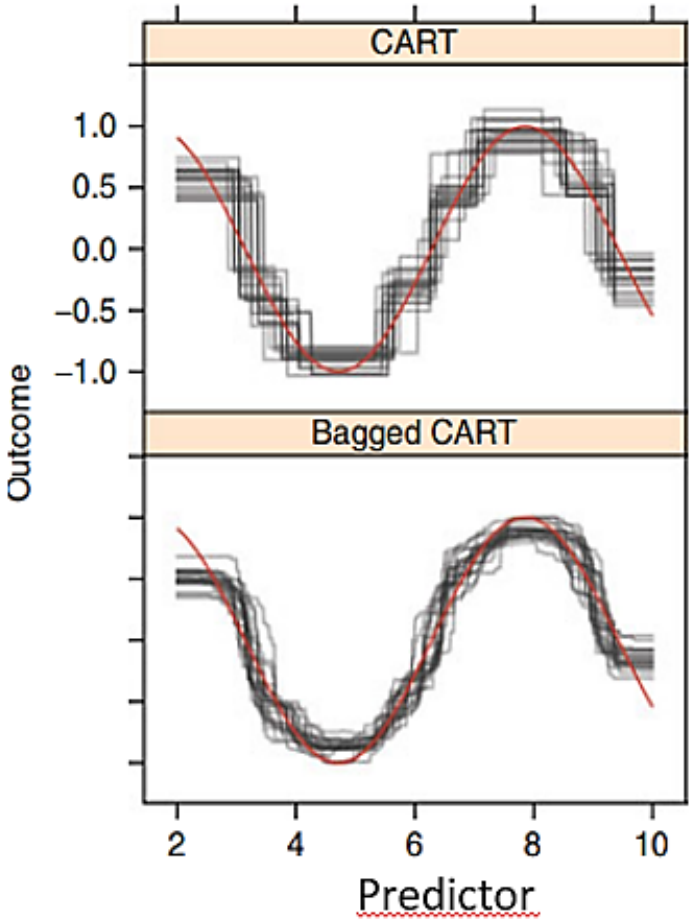


Modele de tip bagging

Forma generală a modelelor de bagging este prezentată mai jos în pseudocod.

1 for i = 1 to m do
| 2 Generarea unui eșantion bootstrap din setul de date folosit la estimare
| 3 Antrenarea unui model de arbore de decizie folosind eșantionul generat în pasul 2
4 end
5 agregarea rezultatelor modelelor de bază și generarea estimării

Comparația rezultatelor folosind modele individuale și modele ensemble de tip bagging arată o reducere semnificativă a variației rezultatelor față de valoarea observată a variabilei dependente (curba roșie).



Kuhn și Johnson (2013)

Algoritmul Random forests

Pornind de la forma generală a modelelor de bagging, cea mai cunoscută implementare este realizată de Breiman (2001) în colaborare cu Adele Cutler.

Algoritmul random forests este prezentat mai jos în pseudocod.

1 Selectarea numărului de modele de bază m
2 for i = 1 to m do
3 Generarea unui eșantion bootstrap din setul de date folosit la estimare
4 Antrenarea unui model de arbore de decizie folosind eșantionul anterior generat
5 for each split (partiționare a setului de date inițial în subdiviziuni) do
6 Selectarea aleatoare a k variabile explicative obținute < P (mulțimea tuturor variabilelor explicative
7 Selectarea variabilei cu cea mai bună putere de predicție și efectuarea splitării
8 end
9 Folosirea unor criterii de stopare a estimării modelului de bază i (fără toaletare)
10 end
11 agregarea rezultatelor modelelor de bază și generarea estimării (inclusiv calculul indicatorilor referitori la performanța modelului).

Kuhn și Johnson (2013)

Modelul Random forests: Caracteristici

- folosește modele de bază cu bias redus și varianță mare în scopul reducerii celei din urmă
- Poate folosi arbori de decizie CART/ Rpart și arbori de decizie bazați pe inferență condițională (cum ar fi C4.5 și C5.0)
- Timpul de estimare mai redus deoarece în estimarea modelelor de bază nu sunt folosite toate variabilele cu potențial de predicție (vezi pasul 6 din pseudocod)
- posibilitatea de paralelizare a estimării, prin estimarea modelelor individuale folosind mai multe resurse informatice
- Relația dintre variabilele predictive și variabila dependentă nu mai poate fi determinată/cuantificată. Însă este posibil calculul indicatorului **variable importance** la nivel de model random forests

Random Forests- Optimizare

Estimarea se poate optimiza (**hyperparameter tuning**) prin ajustarea următorilor parametri (denumiți în practică **hiperparametri**):

- m_{try} numărul variabilelor predictive care trebuie folosite în estimarea fiecărui model de bază. Pentru variabile țintă de tip categorial de regulă m_{try} =P/3, iar pentru cele de tip continuu, m_{try} = $\sqrt{P}$
- numărul de modele de bază folosite (minimum 500, de preferat peste 1000).

Mai există și alte posibilități de optimizare:

- Numărul minim de observații dintr-un nod terminal
- Numărul maxim de noduri terminale

Deși pot fi utile pentru evitarea supraantrenării, cele două opțiuni pot impune restricții nejustificate asupra modelului

Kuhn și Johnson (2013)

Modelul Random forests- evaluarea rezultatelor

Rezultatele obținute în urma estimării pot fi evaluate cu aceleași metode folosite pentru alte modele, în funcție de tipul variabilei dependente (continuă sau discretă).

Structura complexă a estimării îngreunează înțelegerea contribuției/influenței fiecărei variabile dependente la performanța modelului.

Însă modelele de tip bagging sau boosting implementează măsuri de importanță a variabilei (variable importance), similare cu cele implementate pentru modelele de arbori de decizie.

Kuhn și Johnson (2013)

Modelul Random forests- evaluarea rezultatelor (II)

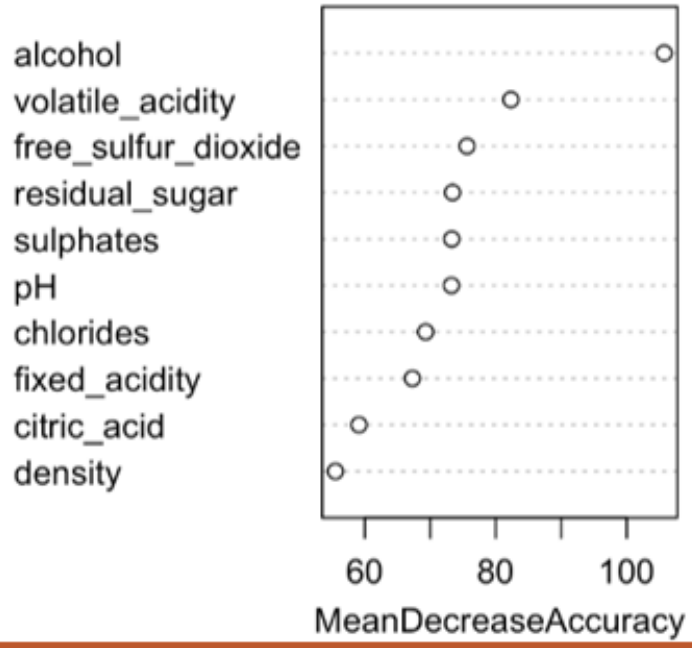
Evaluarea importanței variabilelor diferă în funcție de tipul variabilei dependente.

Pentru variabilele de tip discret, se folosește:

- Descreșterea medie a acurateței modelului datorată scăderii performanței indicatorului care exprimă impuritatea nodului (de exemplu impuritatea de tip Gini pentru arborii de decizie CART/ rpart).

- Descreșterea medie a acurateței modelului în urma reeșantionării aleatoare a variabilei explicative analizate în setul de date out-of-bag (de validare).

	MeanDecreaseGini
AGE	195.757979
GENDER	53.471808
BALANCE	316.586168
OCCUPATION	101.506594
AGE_BKT	102.378902
SCR	312.499265
HOLDING_PERIOD	224.278146
ACC_TYPE	23.599333
ACC_OP_DATE	276.816419
LEN_OF_RLTN_IN_MNTH	252.258727
NO_OF_L_CR_TXNS	202.613633



Modelul Random forests- evaluarea rezultatelor (III)

```
      %IncMSE IncNodePurity
cyl  17.058932     181.70840
disp 19.203139     242.86776
hp   17.708221     191.15919
...
```

Pentru variabilele de tip continuu, se folosește:

- Descreșterea medie a acurateței modelului datorată scăderii performanței indicatorului care exprimă impuritatea nodului (de regulă suma pătratelor erorilor).

- Descreșterea medie a acurateței modelului în urma reeșantionării aleatoare a variabilei explicative analizate în setul de date out-of-bag (de validare). De regulă se folosește mean squared error.

Kuhn și Johnson (2013)

Mulțumesc pentru atenție!