Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Mesterséges Intelligencia és Rendszertervezés Tanszék

VIMIAC16 2025/26/I.

Együttes tanulás

Előadó: Dr. Hullám Gábor

Predikciós teljesítmény javítása 1.

Lehet-e az előrejelzés pontosságát és megbízhatóságát javítani?

A: „egy modellünk van”

- Regularizáció – túltanulás ellen védekezünk

- Hiperparaméter hangolás
  - Valójában több hasonló modellünk van hiperparaméterek szerint
  - Ezek közül választjuk ki a „legjobbat” → modellszelekció

- Keresztvalidációs technikák
  - Valójában több hasonló modellünk van adatpartíciók felhasználása szerint
  - Itt inkább az átlagos teljesítmény, a jósági mutatók tartománya az érdekes →megbízhatóság

Predikciós teljesítmény javítása 2. – Modellek együttese

B: *„több, különböző modellünk van”*

- Lehet-e több modell predikciójának együttesét (pl. átlagát vagy egyéb aggregált eredményét) felhasználni?
- Igen!
- **Együttes tanulásnak** nevezzük azt a folyamatot, amely során adott probléma megoldására több, különböző modellt készítünk, majd ezen modellek predikcióját aggregáljuk a végleges kimenet előállításához.

Modellek együttese

Előny

- Ha több modellt tanítunk, a modellek összessége átlagosan jobb predikciót adhat egy-egy modellnél

Hátrány

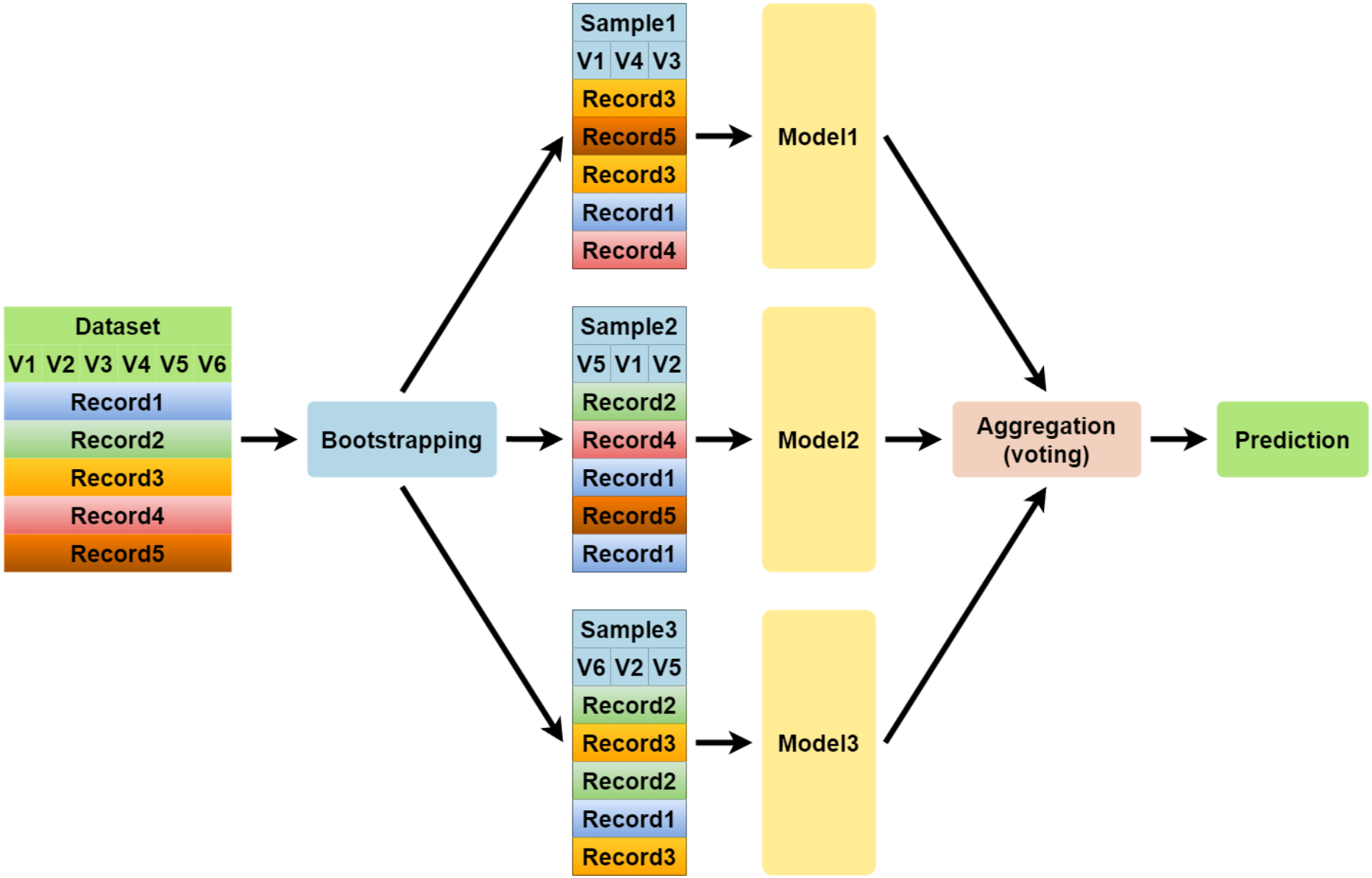
- Több modell több tanítással, nagyobb erőforrásigénnyel jár

Bagging

A bagging lépései:

1. Vegyünk kiindulásul egy N rekordból és K változóból álló adathalmazt.
2. Válasszuk ki egyenletes eloszlás szerint a változók $k \in K$ valódi részhalmazát (ezek a $K*N$ méretű adatmátrix oszlopai), majd visszahelyezéses mintavételezéssel (Bootstrapping) egy N méretű rekordhalmazt a kiválasztott változókra.
3. Illesszünk egy modellt az így kapott adathalmazra.
4. Ismételjük a 2. és 3. lépéseket, amíg egy előre meghatározott számú modellt nem kapunk.

Bagging


Dataset
V1 V2 V3 V4 V5 V6
Record1
Record2
Record3
Record4
Record5
Bootstrapping
Sample1
V1 V4 V3
Record3
Record5
Record3
Record1
Record4
Sample2
V5 V1 V2
Record2
Record4
Record1
Record5
Record1
Sample3
V6 V2 V5
Record2
Record3
Record2
Record1
Record3
Model1
Model2
Model3
Aggregation (voting)
Prediction

Bagging

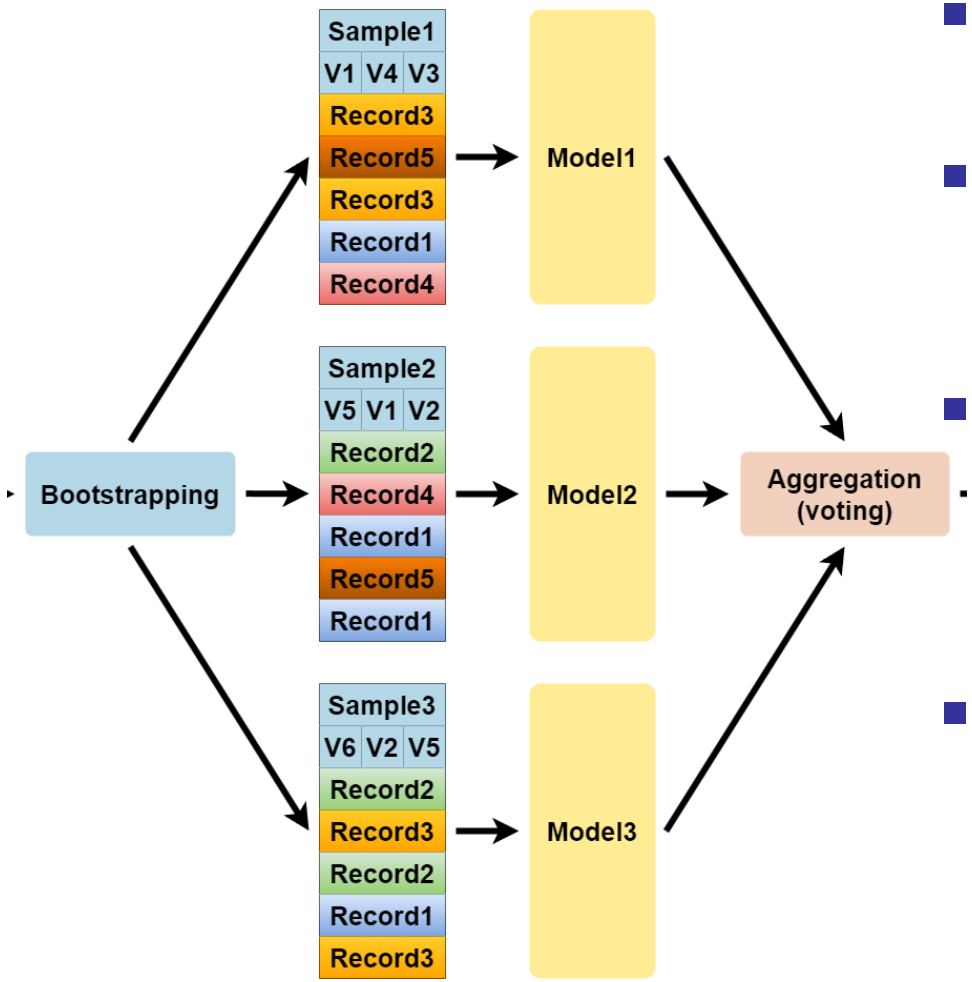
- Ismételjük a 2. és 3. lépéseket, amíg egy előre meghatározott számú modellt nem kapunk.

- 5. Az így kapott modellhalmaz új mintára (tesztre) adott predikcióját aggregáljuk (Aggregation)
  - osztályozás esetén többségi döntéssel,
  - regresszió esetén átlagolással állapítjuk meg.

Heterogén modellek

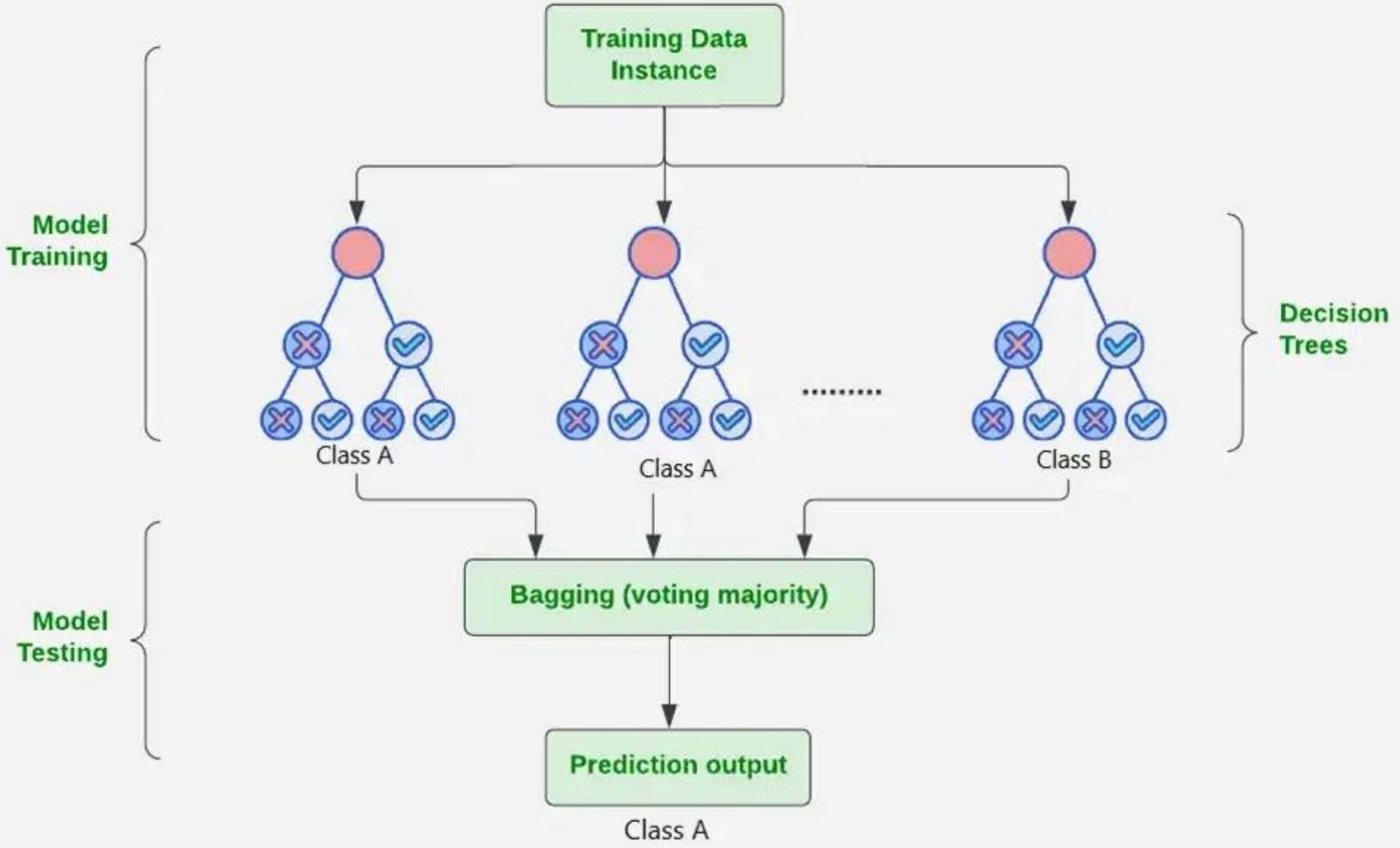
- A 2. lépésben (**Bootstrapping**) mintavételezéssel kialakított adathalmazban a változók és minták csak egy részhalmaza szerepel
- minden modell az eredeti adathalmaznak különböző részhalmazain tanul
- így biztosítható, hogy a determinisztikus (pl. döntési fa) és az alacsony varianciájú (pl. logisztikus regresszió) tanítási folyamattal rendelkező modelltípusok esetén is heterogén modellegyüttest hozzon létre az eljárás,

Bagging – véletlen erdők



- Döntési fák + bagging = véletlen erdő (random forest)
- az eredeti Bagging algoritmus mindig a teljes változóhalmazra végezte a Bootstrap mintavételezést.
- Véletlen erdők esetén viszont előnyös, ha a változók halmaza is véletlenszerűen sorsolt, mivel ekkor az egyes modellek kevésbé korreláltak egymással.
- Ennek a technikának az általános neve *Feature Bagging*, véletlen erdőknél a Bagging részét képzi.

Random Forest Algorithm in Machine Learning



https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

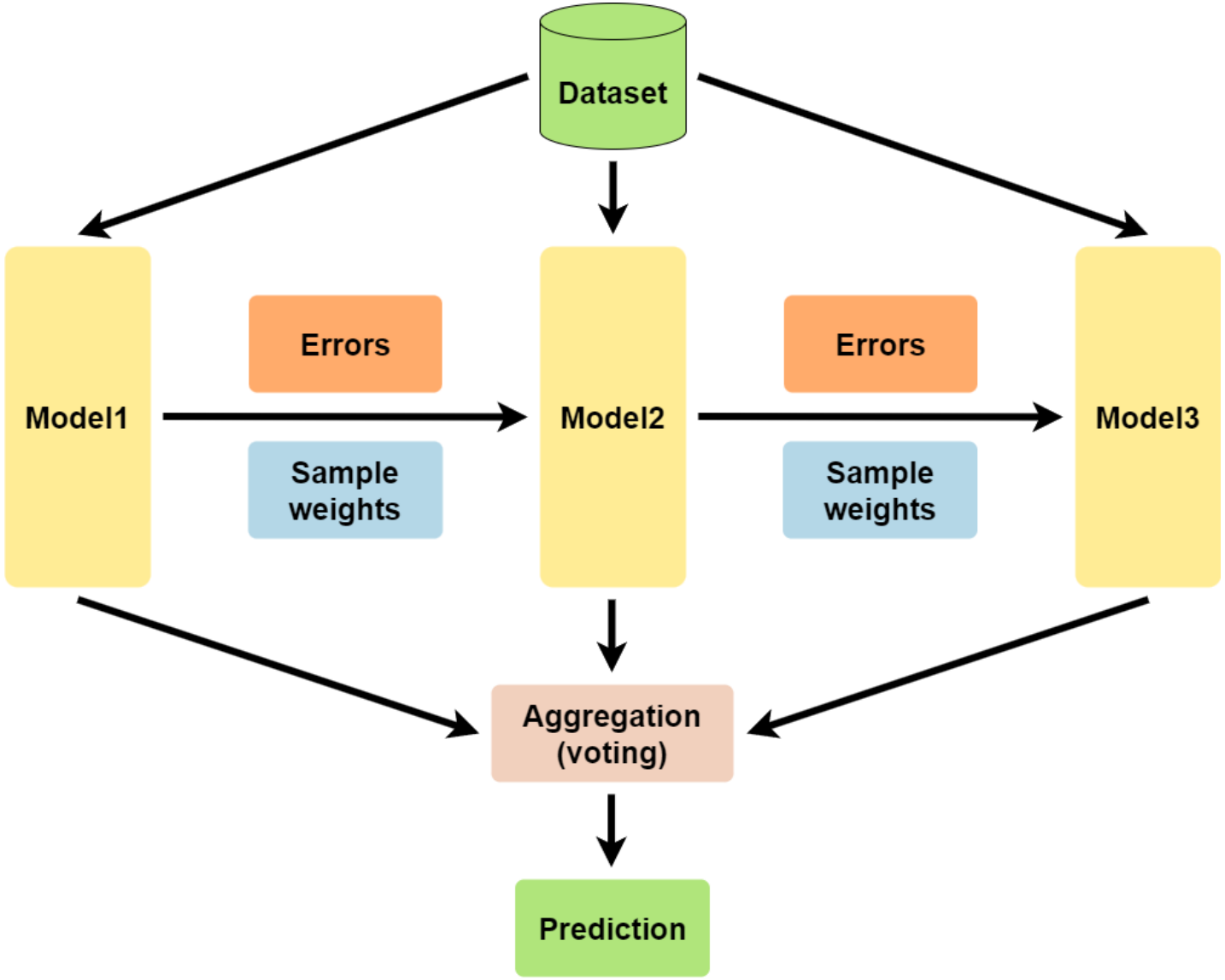
Bagging – előnyök és hátrányok

- + egyes modellek létrehozása párhuzamosítható
  - modellek egymástól teljesen független módon, véletlenszerűen kiválasztott minta alapján lettek tanítva
- - nagy számú modellre van szükség
  - hogy az aggregált predikció pontossága jelentősen jobb legyen az egyes modellekénél.

Boosting

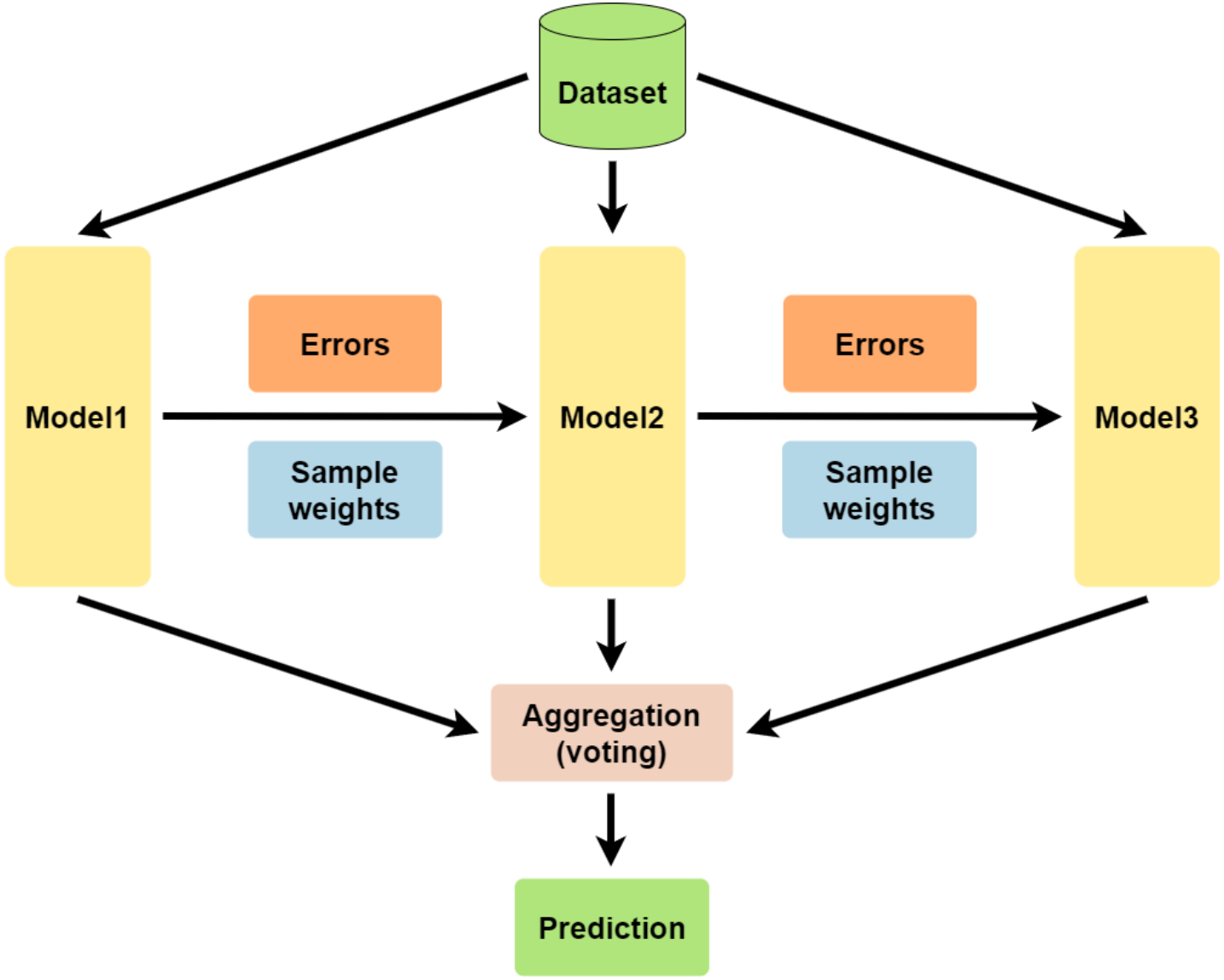
- **Boosting** módszerek szekvenciálisan adnak hozzá újabb modelleket a modellegyütteshez úgy, hogy minden hozzáadott modell valamilyen módon "tanul" a már elkészült modellek gyengeségeiből.
- Az egyik legkorábbi Boosting algoritmus az Adaptive Boosting (röviden **AdaBoost**)
  - minden modell tanítása során nagyobb súlyt kapnak azok a tanítópéldák, amelyekre a korábbi modellek hibás predikciót adtak, ezáltal az újonnan létrehozott modell azokon a mintákon jobban fog teljesíteni.

AdaBoost



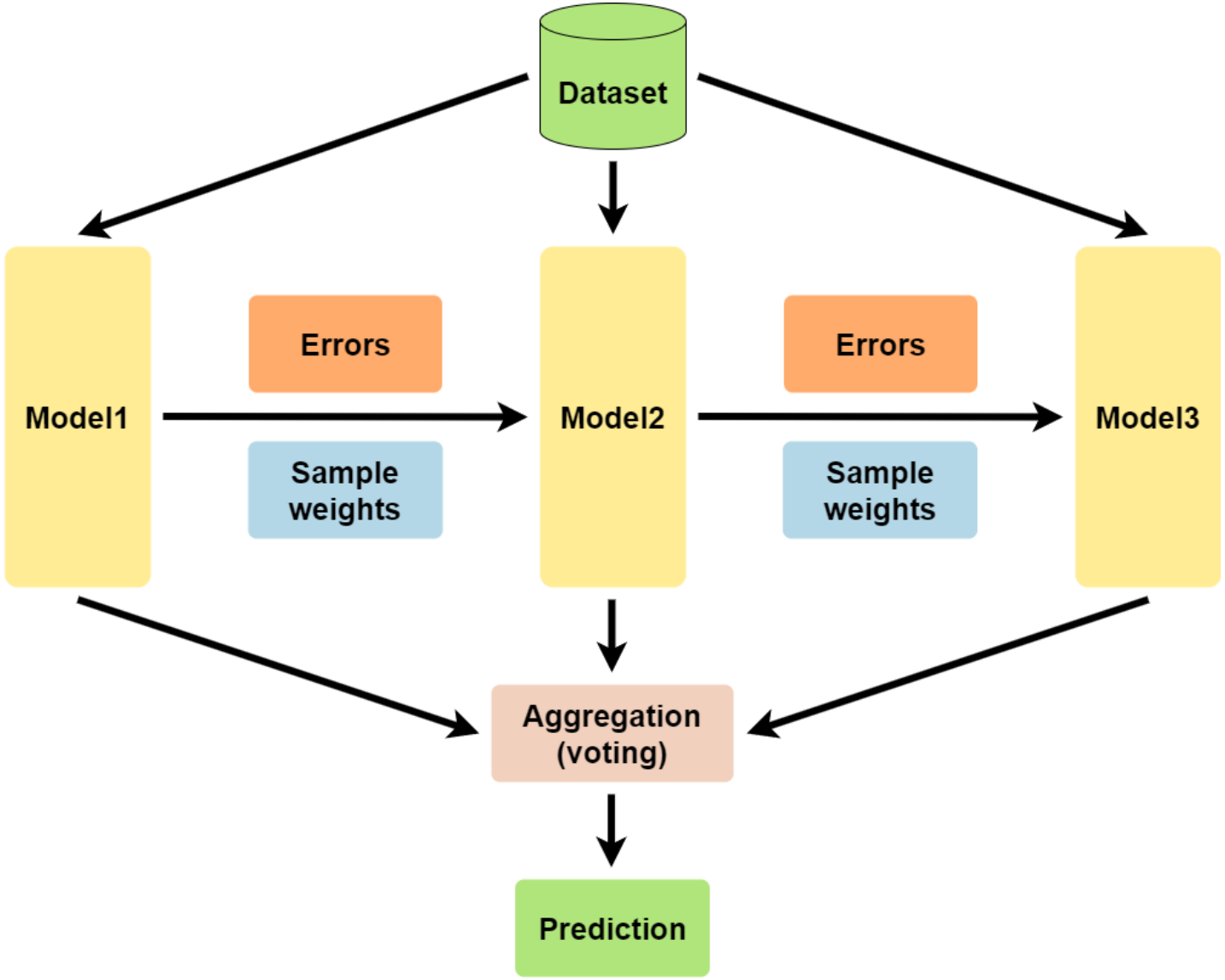
- 1. Kiindulásként határozzunk meg uniform módon egy súlyértéket minden, az adathalmazban található rekordhoz.

AdaBoost



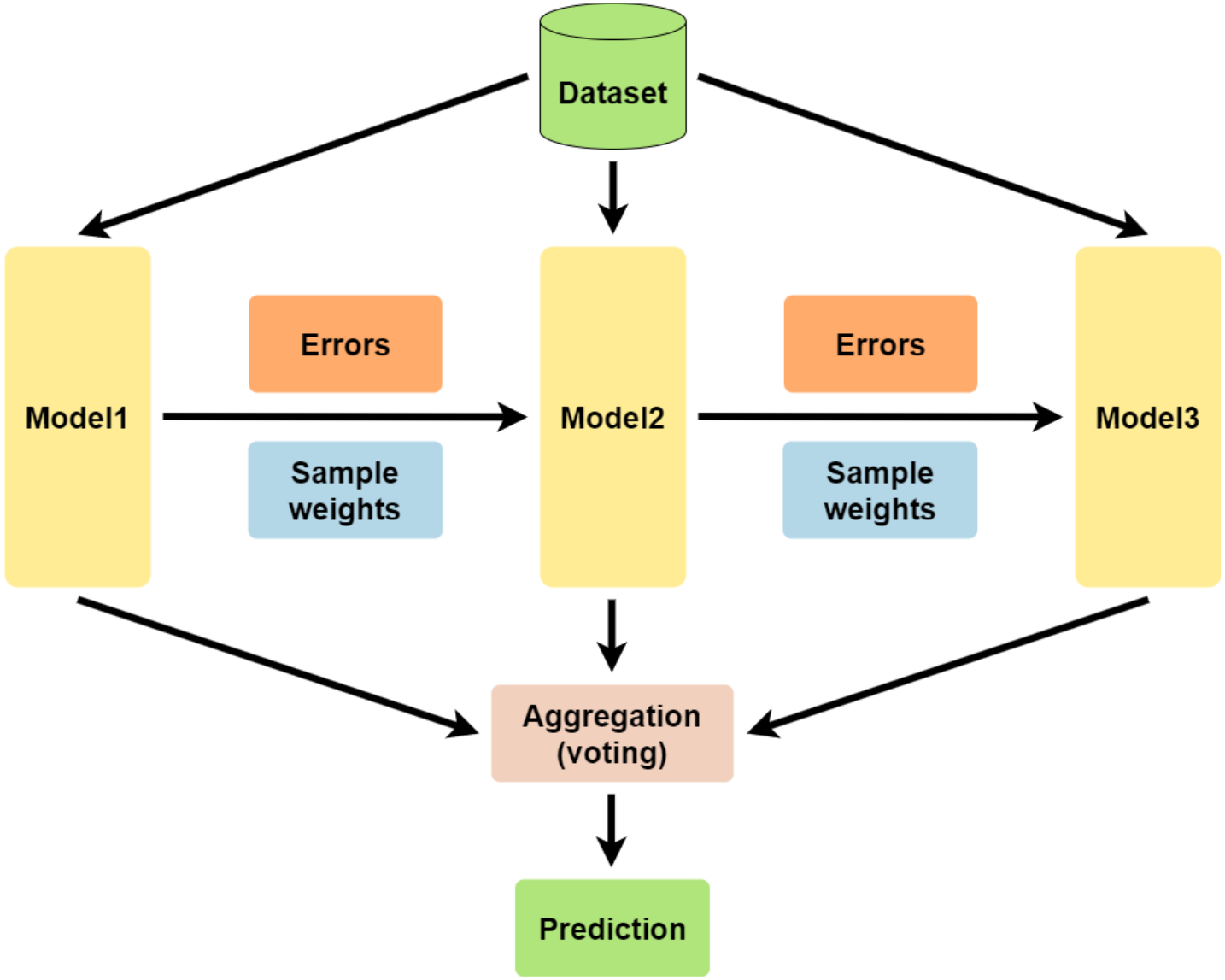
- 2. Tanítsunk be egy modellt a pillanatnyi súlyokkal.
- 3. Növeljük azoknak a mintáknak a súlyát, amelyekre a legújabb modell hibás predikciót adott, majd normáljuk a módosított súlyvektort.

AdaBoost



- 4. Ismételjük a 2. lépéstől, amíg előre megadott számú modellt nem kapunk.
- 5. A modellegyüttes új mintákra adott predikcióját úgy kapjuk, hogy az egyes modellek kimenetét súlyozott módon aggregáljuk.

AdaBoost



- Az aggregáció során az egyes modellek súlya annál nagyobb, minél jobb prediktív teljesítményt adtak a tanítóhalmazon.

Boosting – előnyök és hátrányok

- A Boosting módszerek iteratív módon javítják a meglévő modellhalmaz hibáit
- + gyorsítva a folyamat konvergenciáját
- + hasonlóan jó prediktív teljesítmény eléréséhez kevesebb modell is elegendő, mint Bagging algoritmushoz viszonyítva
- - minden modell tanítása az előző modell hibájától függ
- - egyes modellek létrehozása nem párhuzamosítható, csak szekvenciális módon történhet.

Stacking

- Eddigi feltételezés: a modellegyüttes minden tagja ugyanazon modellcsaládból (pl. döntésifa, SVM, stb...) kerül ki.
- Általánosan: egy modellegyüttes tetszőleges típusú osztályozókból (vagy regresszorokból) is állhat
  - Feltétel: ugyanazon osztályozási (vagy regressziós) problémára kínálnak megoldást.
- Azt az általános esetet, amikor egy modellegyüttes eltérő modellcsaládból származó becslőkből tevődik össze, **Stacking**-nek nevezzük.

Stacking

- Előnyt jelenthet a modellek diverzitása
- az eltérő megközelítést képviselő modellek predikciói jól kiegészíthetik egymást
- a modellegyüttesben lévő egyes modellek jellemzően eltérő bemeneteken tévednek, így az általuk hozott egyszerű többségi döntés megbízhatóbb, mint az egyes modellek predikciója.
- ez nem feltétlenül igaz minden osztályozási (vagy éppen regressziós) problémára

Stacking

- a különböző megközelítések egyéni vizsgálata továbbra is fontos
- előfordulhat, hogy a modellek többsége különösen rosszul teljesít az adathalmazon, így a modellegyüttes teljesítménye akár rosszabb is lehet néhány egyéni modell teljesítményénél.