Funkcija gubitka (*loss*)

$$L\left(y, h_{\hat{\theta}}(x)\right)$$

Koliko gubimo ako koristimo ovaj model umesto savršenog modela?

- Savršene predikcije → gubitak = 0

L1: $L\left(y, h_{\hat{\theta}}(x)\right) = \left|y - h_{\hat{\theta}}(x)\right|$

L2: $L\left(y, h_{\hat{\theta}}(x)\right) = \left(y - h_{\hat{\theta}}(x)\right)^2$

Ovo su primeri *simetričnih* funkcija gubitka

Funkcija gubitka (*loss*)

$$L\left(y, h_{\hat{\theta}}(x)\right) = RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - h_{\hat{\theta}}(x^{(i)})\right)^2}$$

Root MSE
(ista merna jedinica kao y)

Funkcija fitovana korišćenjem *trening* podataka

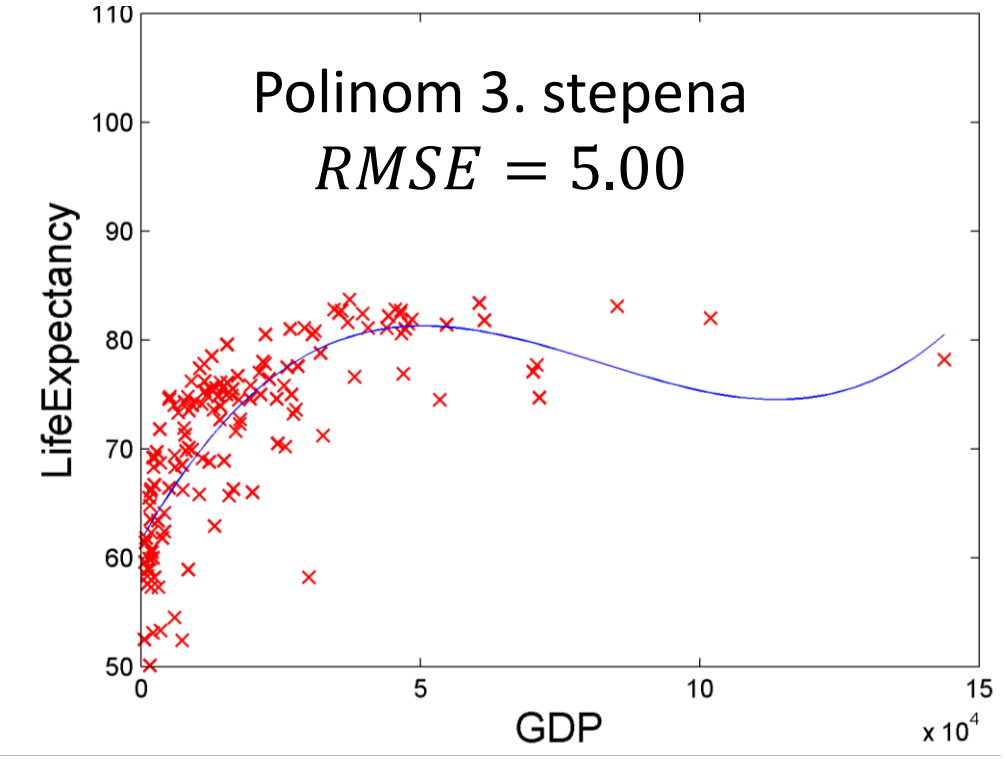
Šta je cilj mašinskog učenja?

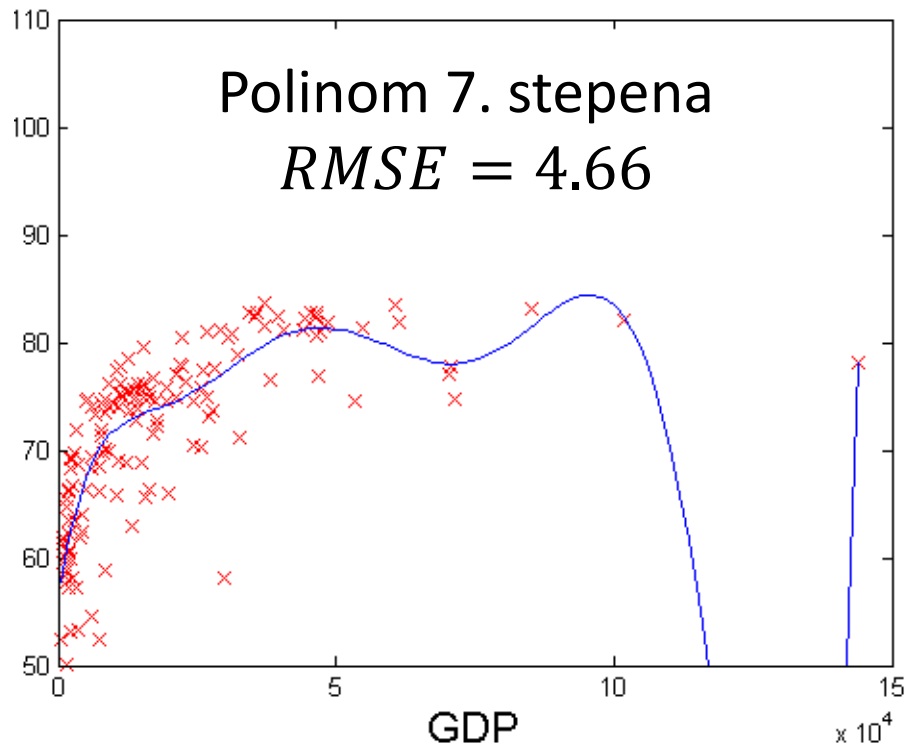
Dobra generalizacija

Da dobro određujemo vrednost izlaza za *nove* primere

(različite od onih iz trening skupa)

Gubitak na TRENING skupu





Greška je previše optimistična:
$\hat{\theta}$ je prilagođen baš podacima na kojima merimo gubitak

Ne znamo šta će se zaista desiti na novim podacima

Šta možemo?

Možemo da je procenimo

Trebaju nam primeri koji se ne nalaze u trening skupu

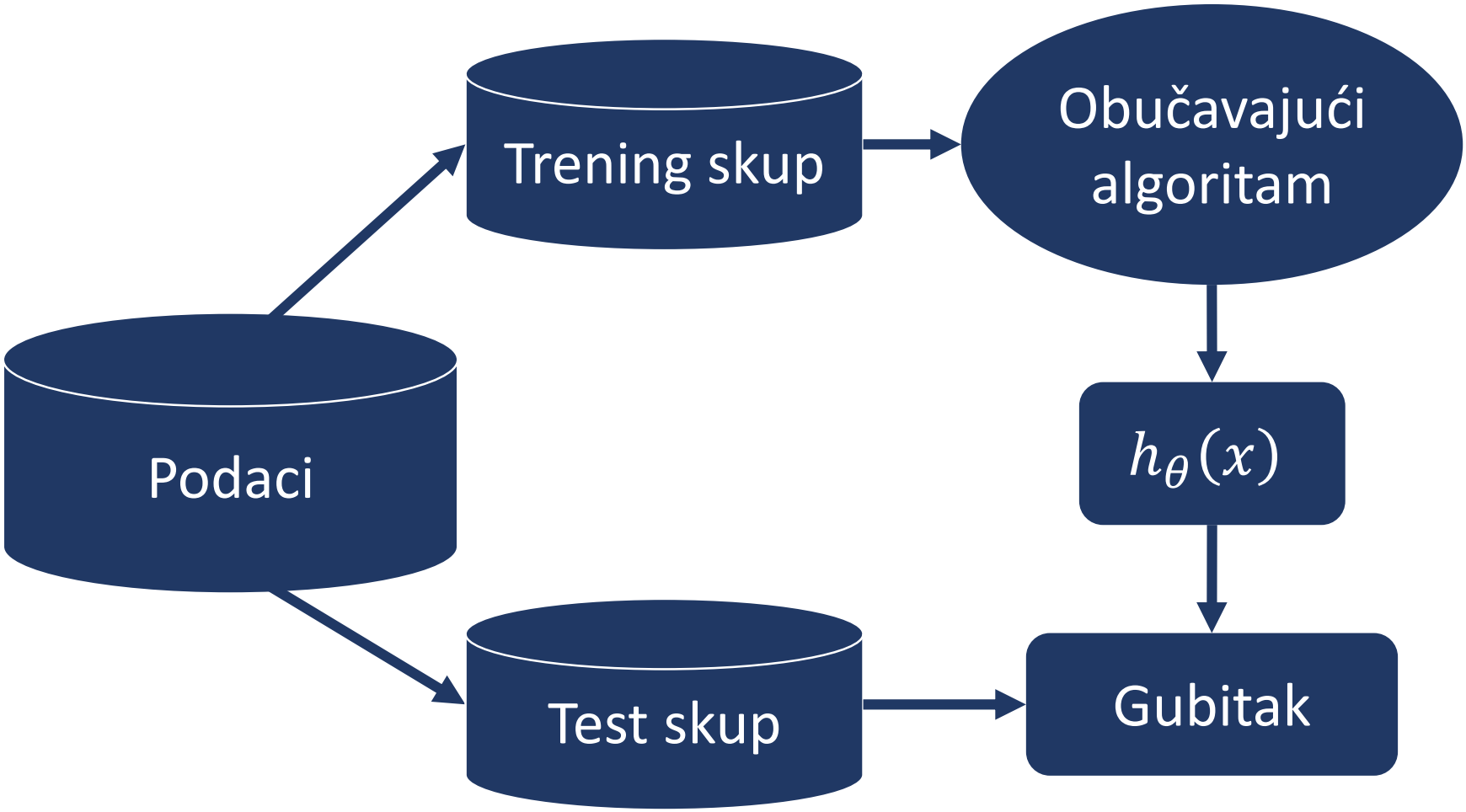
Šta želimo?

Generalizacionu (pravu) grešku: gubitak na SVIM mogućim primerima

SVI mogući odnosi GDP i životnog veka

Problem: ne možemo je odrediti

Postupak obučavanja i evaluacije



Kako podeliti podatke?

Tipično: 90/10, 80/20, 70/30

Ne zaboravite da izmešate podatke

Premalo za dobru procenu θ

Premalo za dobru procenu generalizacione greške

Stratifikacija

- Prilikom deljenja podataka, obezbediti da delovi imaju istu raspodelu kao i ceo skup podataka
- Teško za male skupove podataka
- Pojednostavljena varijanta: očuvati raspodelu ciljne promenljive
  - Sortirati podatke u odnosu na ciljnu promenljivu
  - Ako je *K* broj delova, neka instance sa indeksima $i + j \cdot K$ čine deo P_i za $i = 1, \ldots, K$ i $j = 0,1, \ldots$

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Greška na test skupu (ono šta imamo)

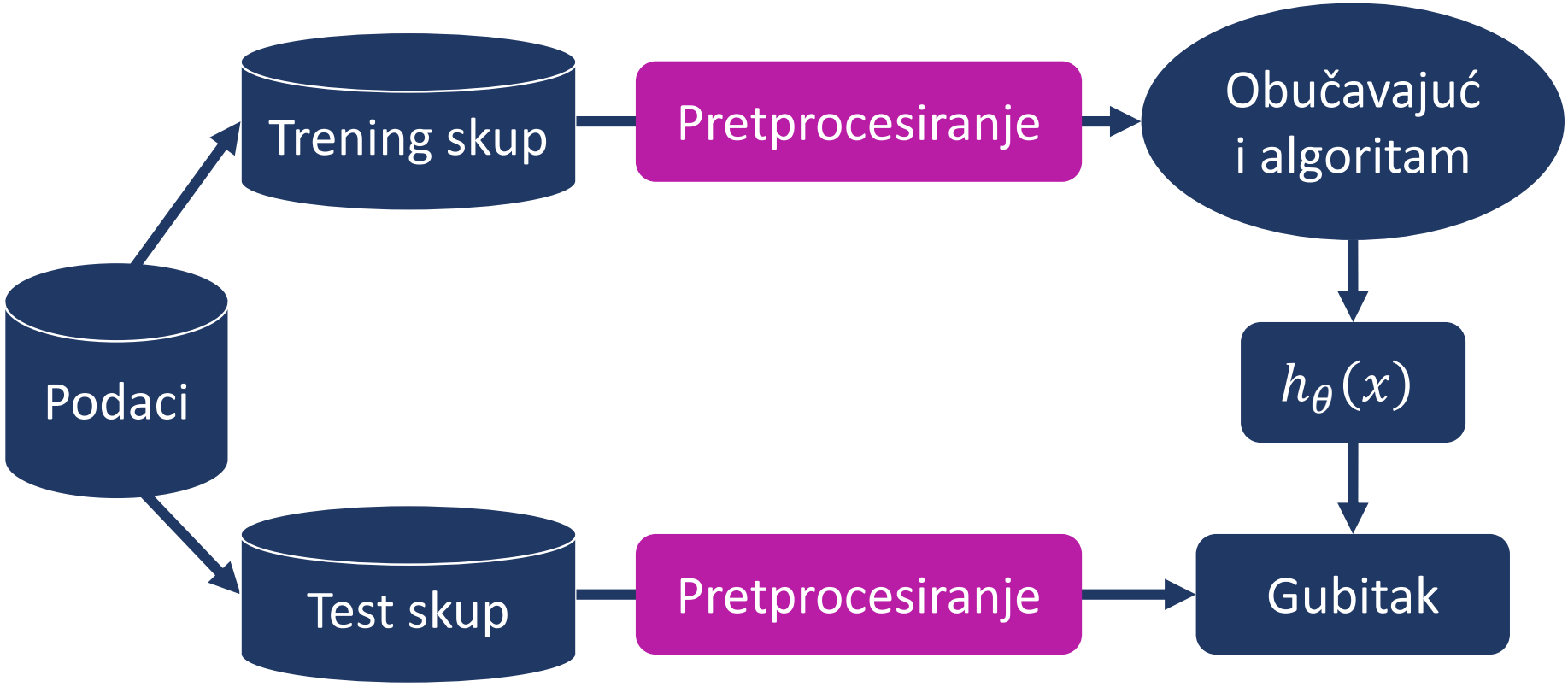
- Možemo je proceniti pomoću primera koji se ne nalaze u okviru trening skupa
  - Podelićemo skup podataka na *trening* i *test* skup
  - Model ćemo trenirati na *trening* skupu
  - Trenirani model ćemo evaluirati na *test* skupu

- RMSE na primerima iz *test* skupa:

$$\frac{1}{N_{test}}\sqrt{\sum_{i=1}^{N_{test}}\left(y^{(i)} - h_{\widehat{\theta}}(x^{(i)})\right)^2}$$

Funkcija fitovana korišćenjem *trening* podataka – **nikada nije videla primere iz *test* skupa**

Postupak obučavanja i evaluacije



Nad trening i test podacima moramo sprovesti istu transformaciju

Mean normalization:
μ_{train} i σ_{train} primenjujemo i prilikom transformacije test podataka