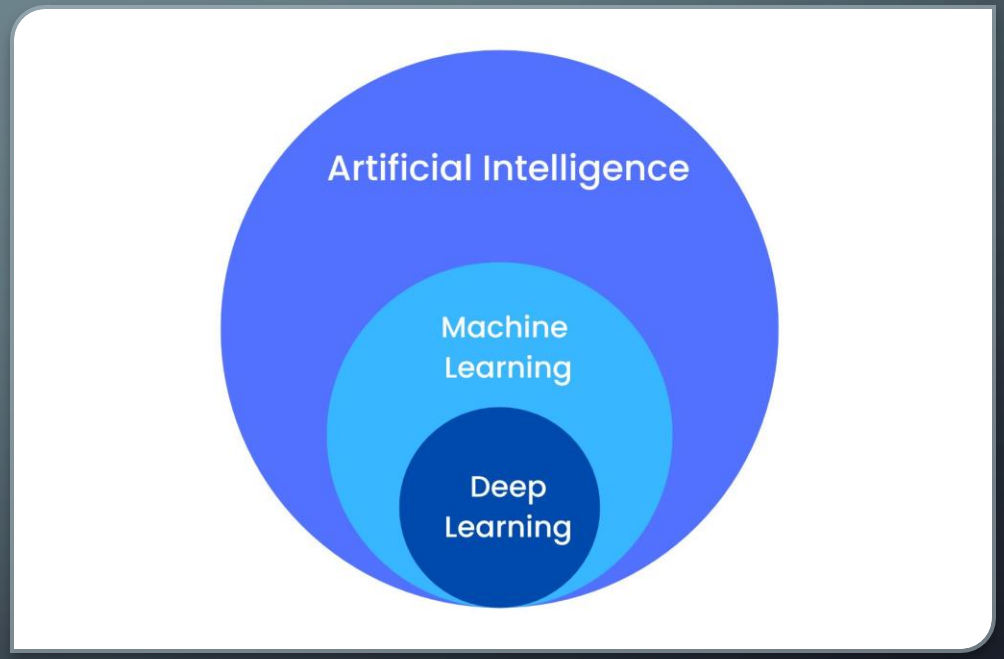
STROJNO UČENJE

V BIOFIZIKI

Simon Perovnik

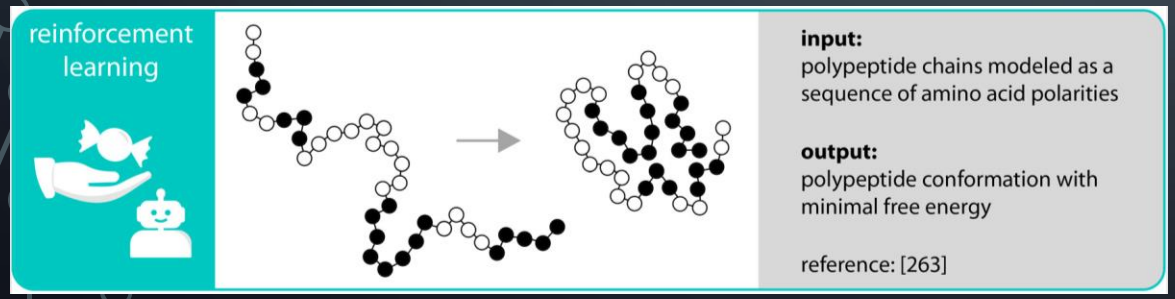
KAJ JE STROJNO UČENJE?

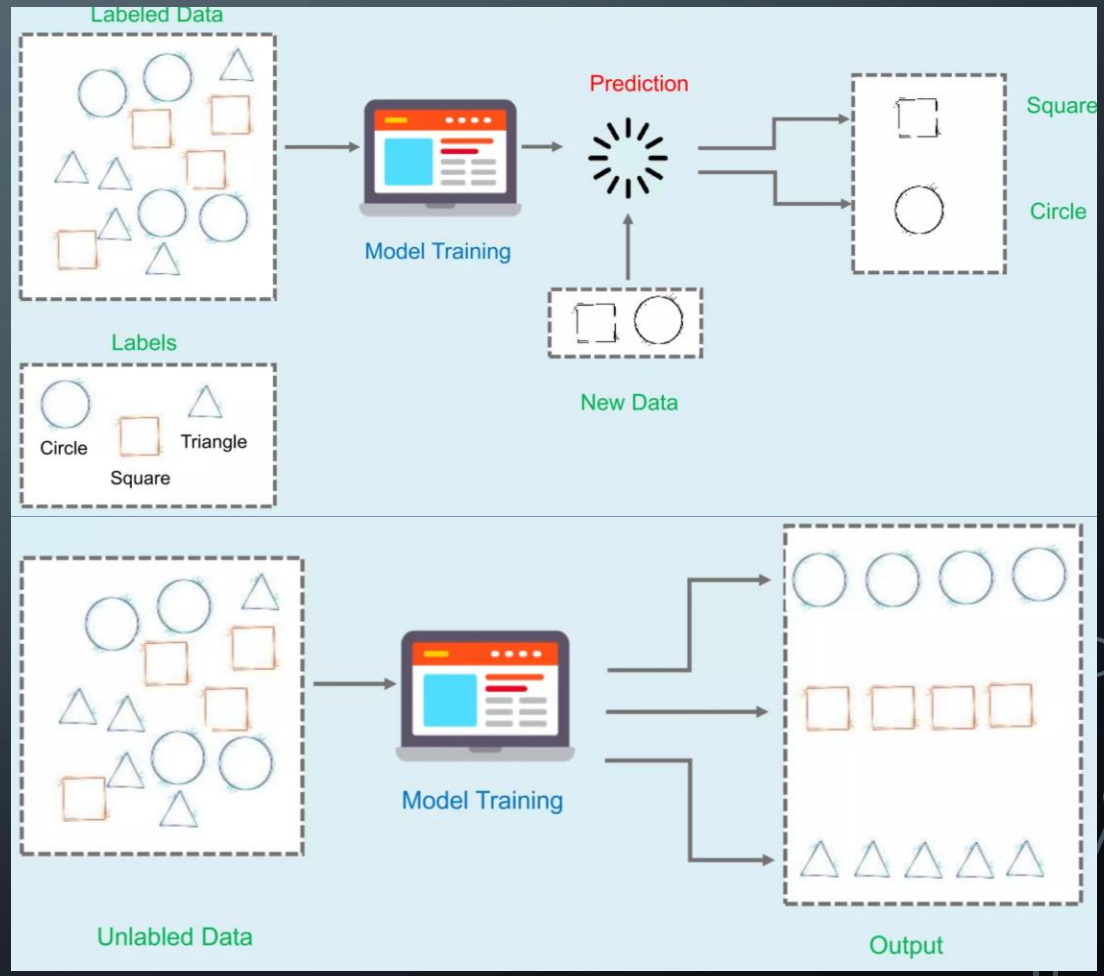
- Učenje je vsak proces pri katerem sistem **izboljša** svoje delovanje preko novih **izkušenj**
- Strojno učenje se ukvarja z računalniškimi programi, ki se avtomatsko izboljšujejo preko novih izkušenj
- Podvrsta umetne inteligence, pri kateri se računalnik "sam" nekaj nauči, ne da bi to **eksplicitno** sprogramirali.
- Glavna ideja: napoved na podobnih podatkih



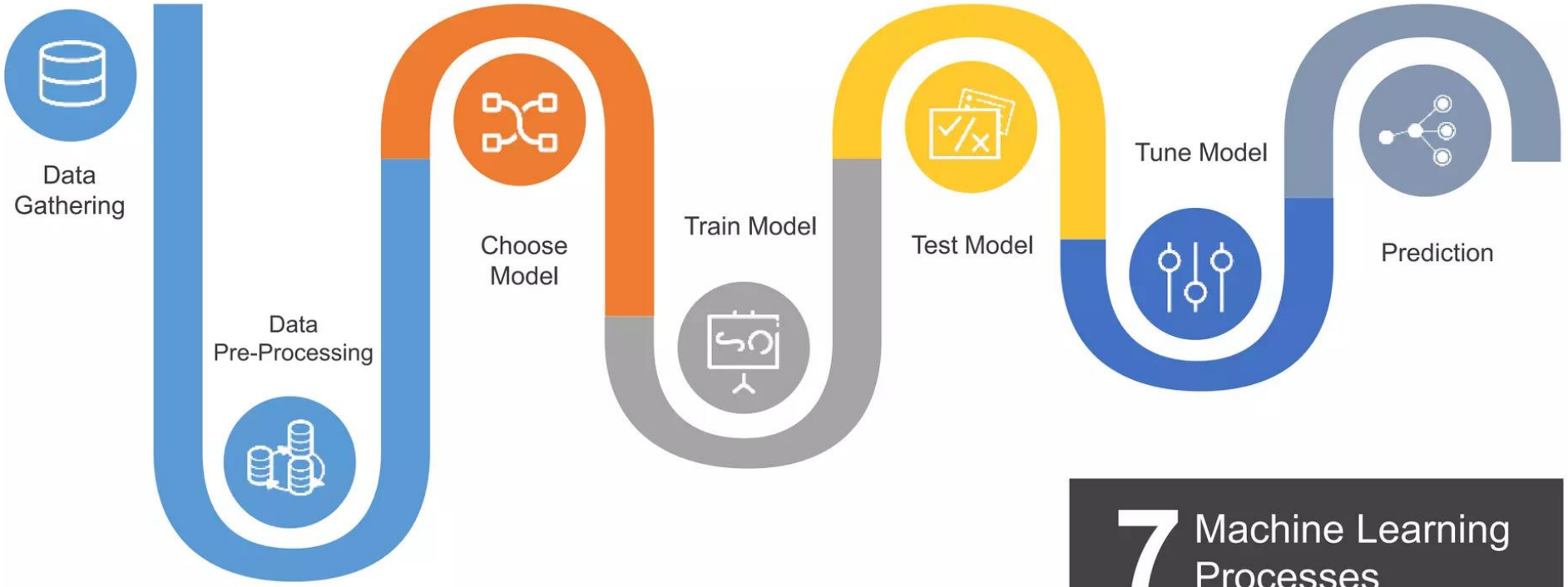
TIPI STROJNEGA UČENJA

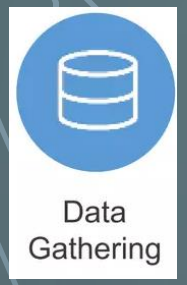
- Nadzorovano učenje
  - Označeni podatki (labeled data)
- Nenadzorovano učenje
  - Neoznačeni podatki
- Spodbujevalno (reinforcement) učenje





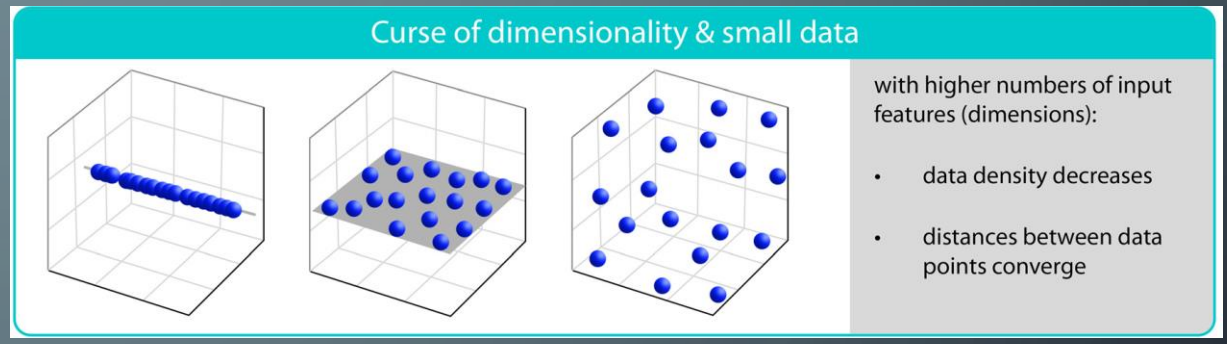
Processes involved in Machine Learning

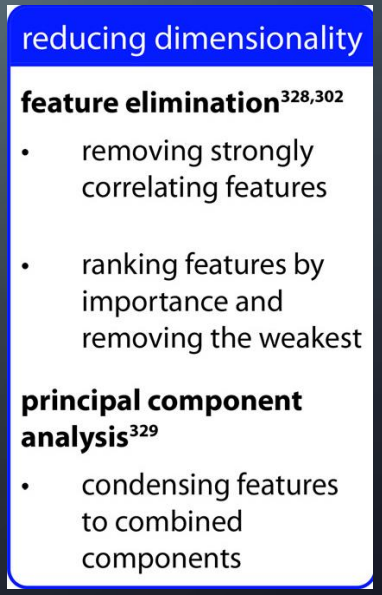


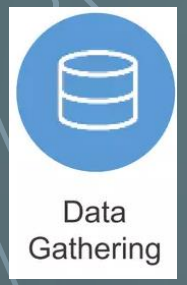


PODATKOVNI SETI

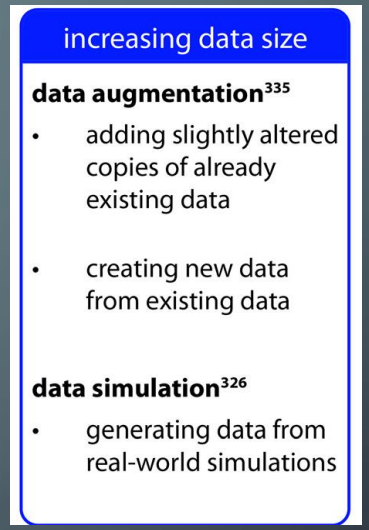
- *Garbage in, garbage out*
- Značilke (features) - fizikalne opazljivke
  - *Curse of dimensionality*
  - Korelirane opazljivke
- Potrebujemo velike sete podatkov (statistični modeli)
  - Označevanje podatkov
- Set za trening in set za testiranje



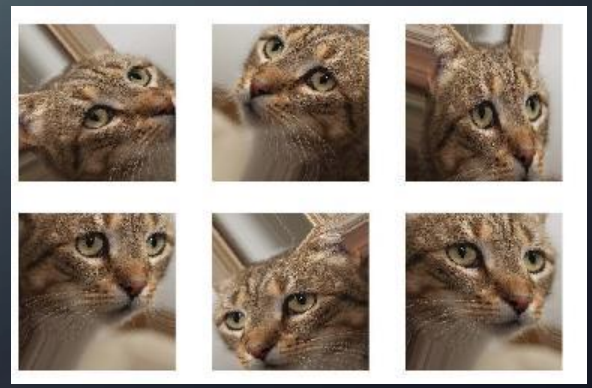


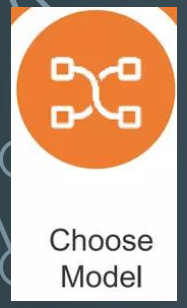


PODATKOVNI SETI



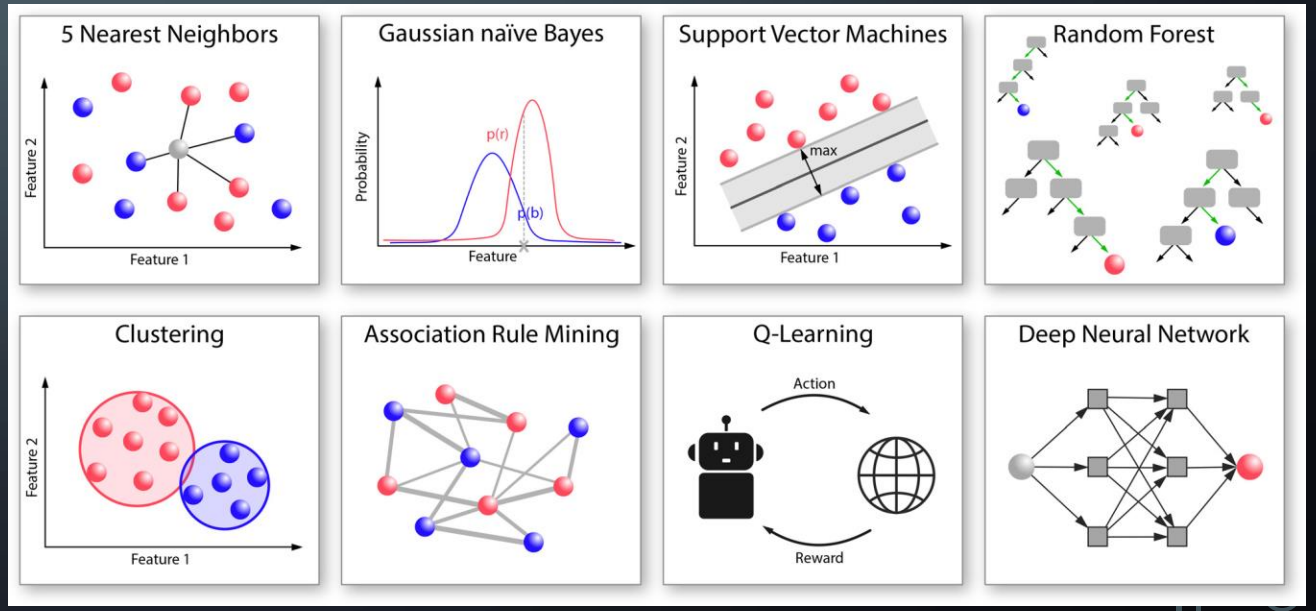
- Umetno povečevanje setov (*data augmentation*)
- Eksperimentalni podatki ali simulacije
- Vektorizacija podatkov
  - Pogosto potrebna pred-obdelava, npr. skaliranje podatkov, maskiranje, …

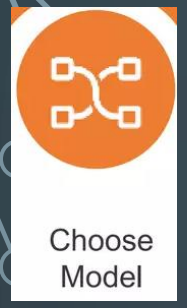




MODELI STROJNEGA UČENJA

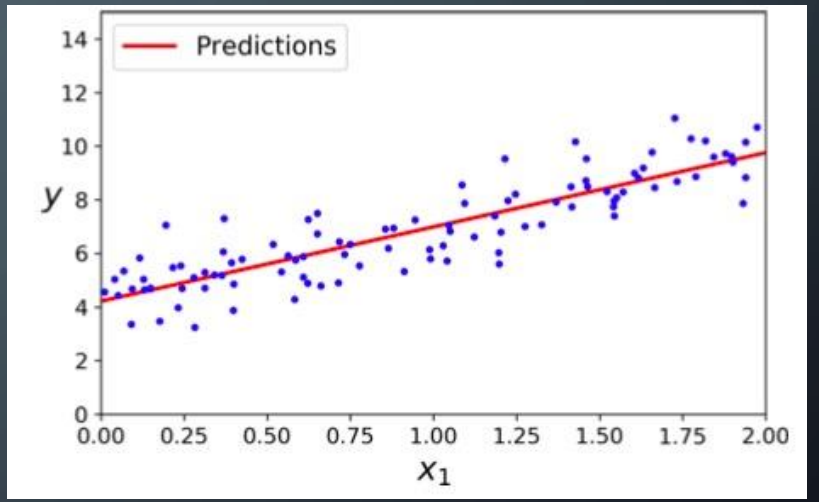
- No free lunch theorem (1996)
  - Če ni nobenih predpostavk (omejitev) na vrsti podatkov, ni formalnega razloga, da bi bil kakšen ML pristop boljši od drugega!
  - Ena metoda ni najboljša za vse probleme
- Regresija, klasifikacija
  - Napoved vrednosti ali razreda

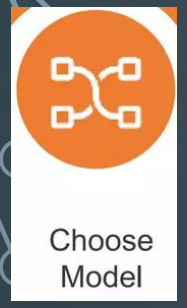




"FITANJE" (REGRESIJA)

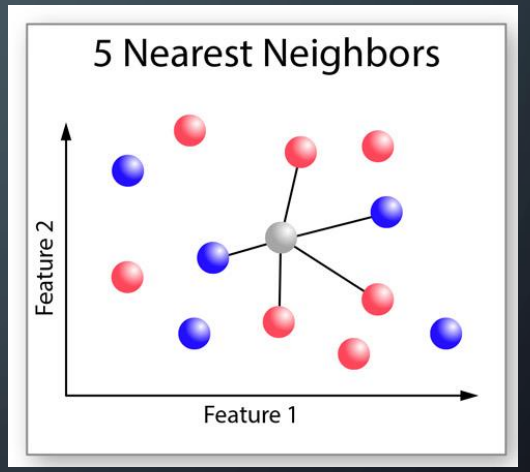
- Model se "uči" z dodajanjem novih podatkov
- Obstaja celo točna rešitev!
- Če imamo veliko količino podatkov - numerika

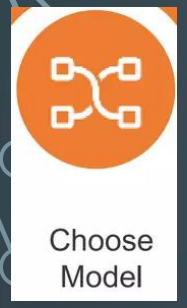




K NEAREST NEIGHBOUR (KNN)

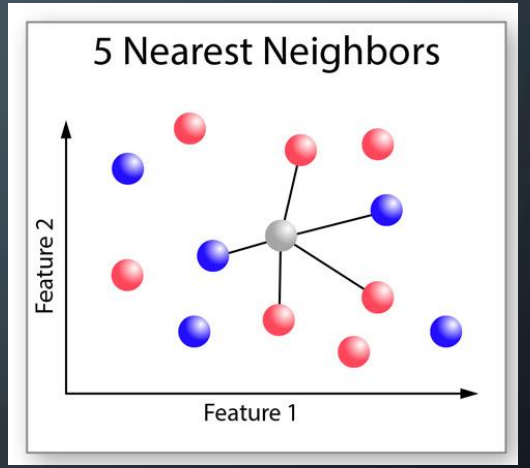
- Klasifikacijski algoritem
- Podatke gledamo v hiperprostoru značilk
- Premisa: Podatki, ki so blizu v hiperprostoru pripadajo istemu razredu
- Glede na že označene podatke za nov podatek:
  - Poiščemo K najbližnjih sosedov
  - Pogledamo katera kategorija je najpogostejša

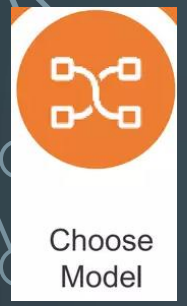




K NEAREST NEIGHBOUR (KNN)

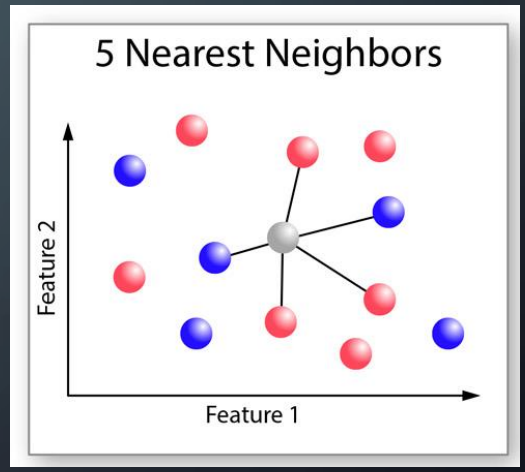
- Metrika razdalje
- Problem *outlierjev* (zahtevajo velik k), a na ta način izgubimo na natančnosti deljenja (robne točke razredov so problematične)
- Uravnoteženost podatkov (predprocesiranje - skaliranje velikosti)
- Curse of dimensionality



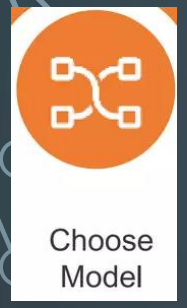


K NEAREST NEIGHBOUR (KNN)

- Primer uporabe:
  - Klasifikacija okuženih s COVID-19: uporaba rentgenskih slik pljuč -> procesiranje slik -> generiranje značilk -> klasifikacije s KNN

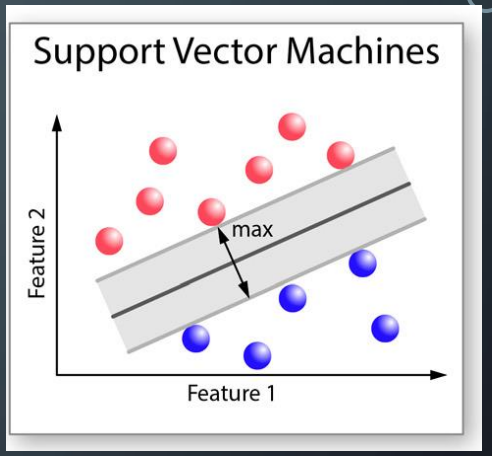


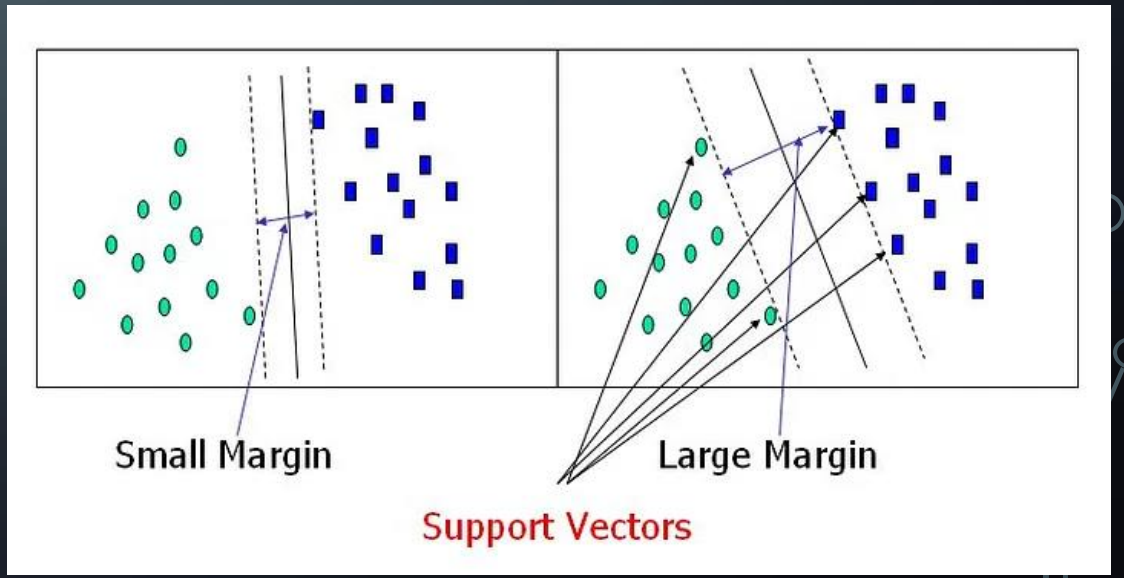
K nearest neighbors[69–72]	
No training phase needed	High dimensionality leads to decreased accuracies
Intuitive and simple algorithm	Can become slow for big datasets
Easily adapts to new training data	Needs feature scaling
Only one hyperparameter to tune	Has problems with imbalanced datasets
	Missing values are problematic

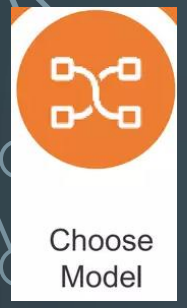


SUPPORT VECTOR MACHINES (SVM)

- Binarna klasifikacija
- V hiperprostoru značilk iščemo hiperpovršino, ki razdeli prostor na dva dela – SVM poišče ploskev, kjer je najširši pas nezasedenosti
  - Potem klasifikacija glede na to, na katerem delu pasu se nahaja

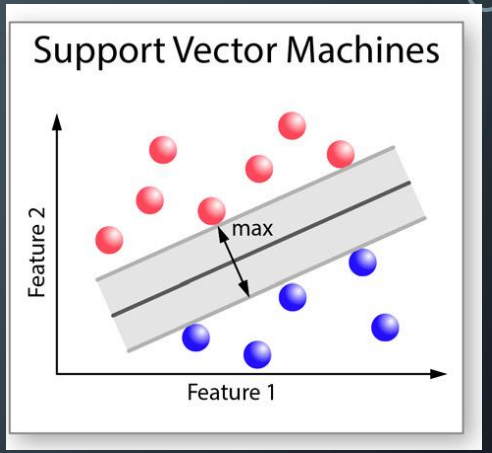


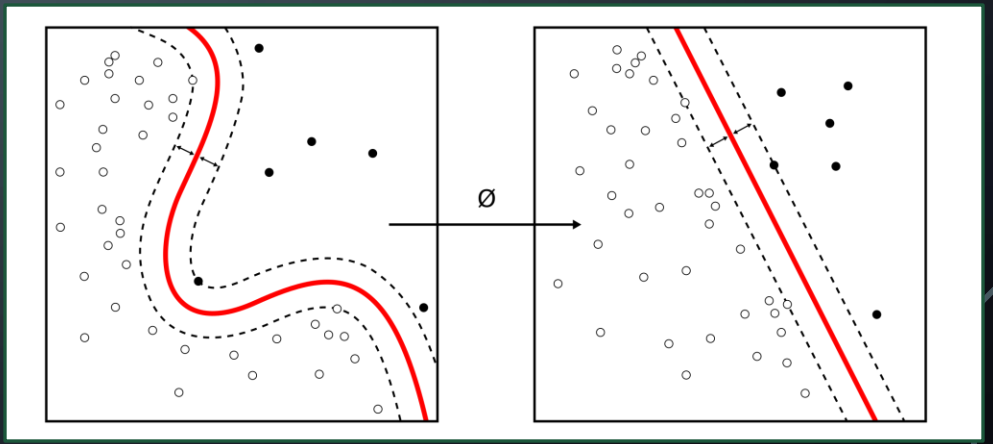


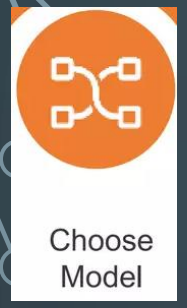


SUPPORT VECTOR MACHINES (SVM)

- Problemi, ko se območja prekrivajo (ali ko so podatki zelo zašumljeni)
- Z uporabo jeder lahko transformiramo podatke razdelimo podatke nelinearno



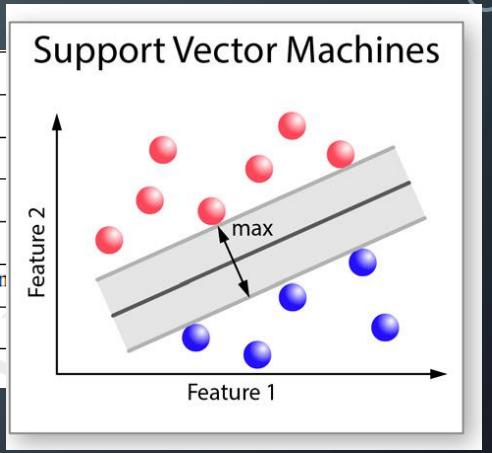


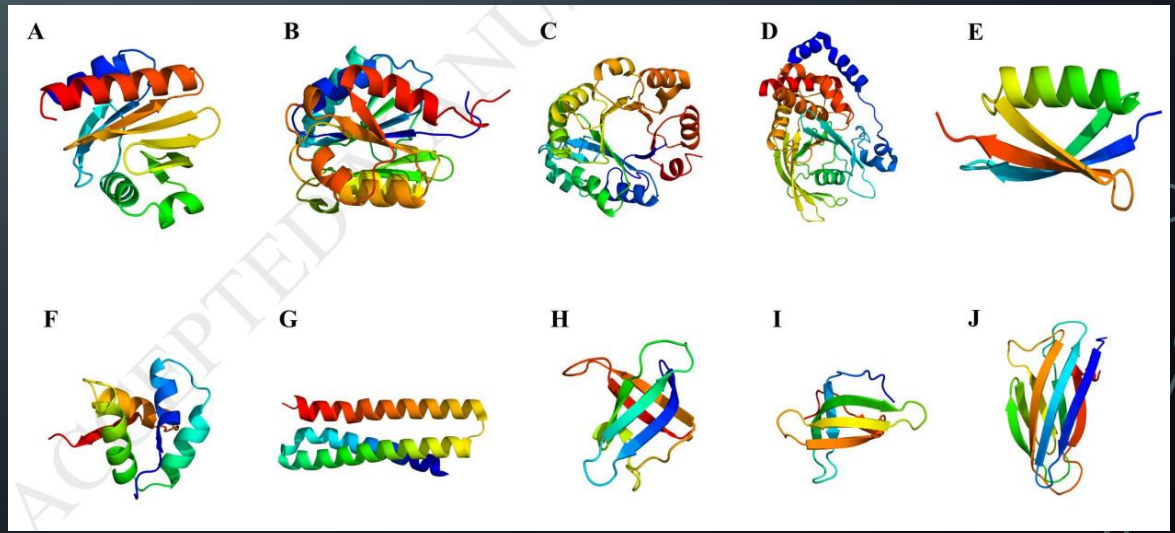


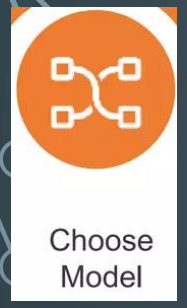
SUPPORT VECTOR MACHINES (SVM)

- Primer uporabe: Analiza strukture zgibanja proteinov
  - Povezava med biofizikalnimi lastnostmi in obliko proteina – proteini, ki imajo različno sekvenco AK, a podobne povprečne biofizikalne lastnosti, se podobno zgibajo
  - SVM uporabljen za klasifikacijo med razredi tipičnih oblik proteinov

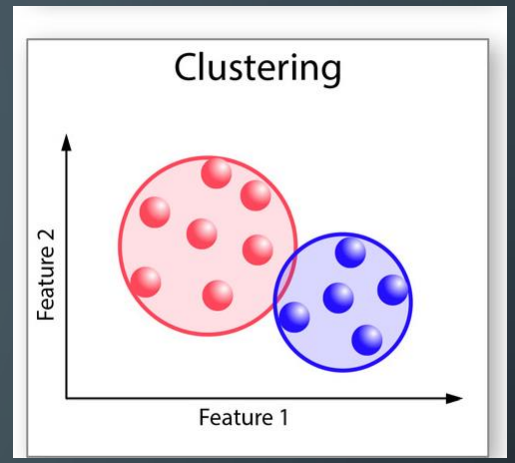
1	**K0**	Compressibility
2	**Ht**	Thermodynamic transfer hydrophobicity
3	**Hp**	Surrounding hydrophobicity
4	**P**	Polarity
5	**pHi**	Isoelectric point
6	**pK'**	Equilibrium constant with reference to the ionization
7	**Mw**	Molecular weight
8	**Bl**	Bulkiness



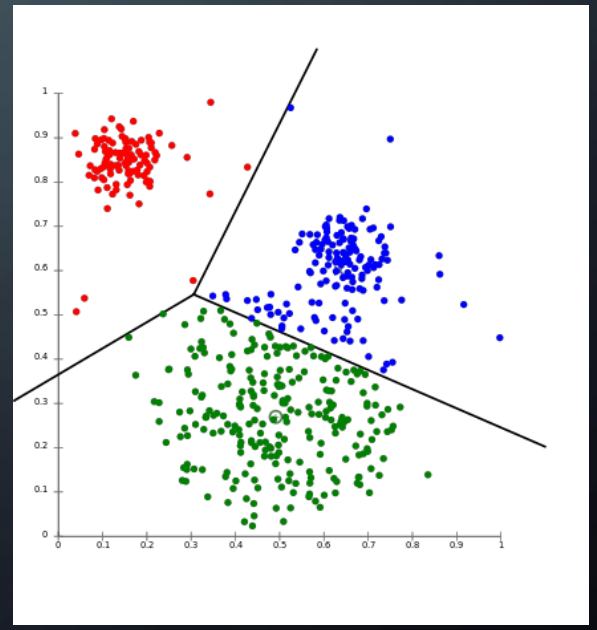


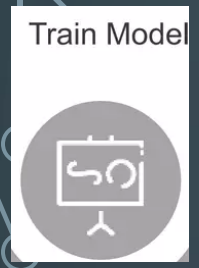


CLUSTERING



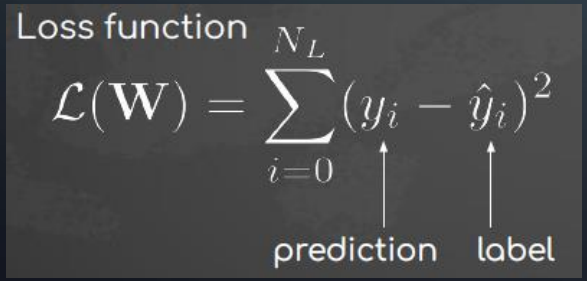
- Nenadzorovana klasifikacija
- K-means gručenje
  - Razdelitev v K gruč preko iterativnega postopka
  - Težave pri različnih velikostih gruč (težko pravilno najti manjše gruče)
  - Kakšen K je pravi?
- Problem računanja razdalj v veliko dimenzijah - drago

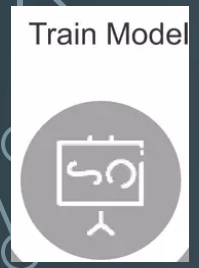




TRENIRANJE MODELA

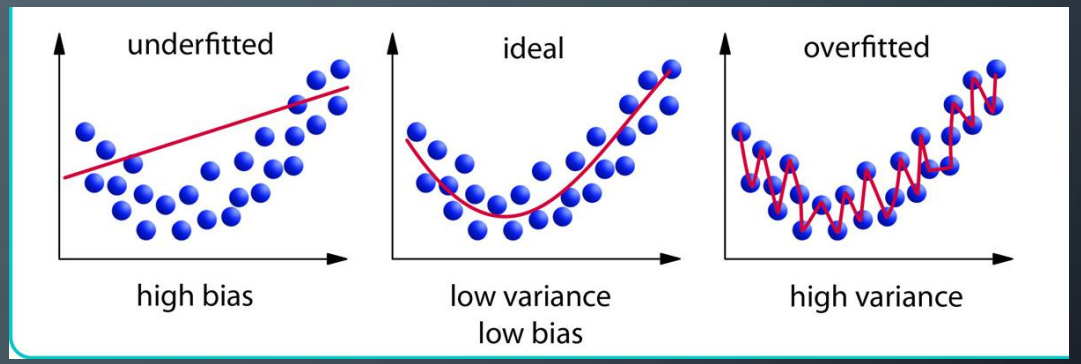
- Matematičen postopek – navadno neke vrste minimizacija
- Definira se Cost/Loss funkcijo
- Iskanje pravih parametrov modela, tako da se minimizira loss funkcijo
  - Linearna regresija – iskanje parametrov linearne funkcije, tako da je npr. MSE minimalen
  - SVM – iskanje parametrov hiperploskve, tako da je pas, ki ločuje razrede najširši
  - Nevronske mreže - iskanje vrednosti uteži, da je output kar se da podoben predvideni oznaki





TRENIRANJE MODELA

- Over-fitting, under-fitting
  - Problem: slaba napovedna moč
  - V več dimenzijah lahko ni očitno, da smo pretirano prilagodili model
  - Rešitve:
    - Uporabimo prav red polinoma za fitanje
    - Early stopping



avoiding overfitting

regularization[330]
- constraining model coefficients

cross-validation[331]
- holding back data to test the model on truly unseen data

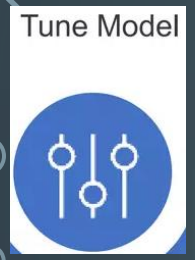
dropout[332]
- ignoring a subset of neurons with a set probability

ensembling[333]
- combining predictions from multiple separate models
- randomly creating subsets of samples by bootstrap aggregation

early stopping[334]
- reducing the number of iterations performed to train a model

TESTIRANJE MODELA

- Napovedujemo na še ne-videnih podatkih in gledamo natančnost
  - Več različnih mer za natančnost, odvisno od konteksta (npr. pri iskanju bolnih oseb, si pred vsem želimo, da ne zgrešimo koga, ki je bolan)
- Testni set podatkov (npr. 20% celotnega seta)
- *Confusion matrix*
- Primerjava z drugimi modeli

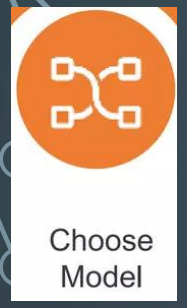


PRILAGANJE MODELA

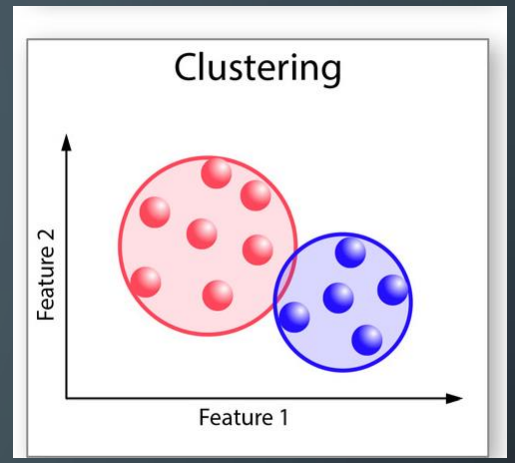
- Spreminjanje hiperparametrov
  - Število gruč, arhitektura nevronske mreže, število iteracij, metrika razdalje, inicializiacija modela
  - Grid-search
- Spreminjanje vhodnih podatkov - dodatna obdelava, drugačna reprezentacija

NAPOVEDOVANJE

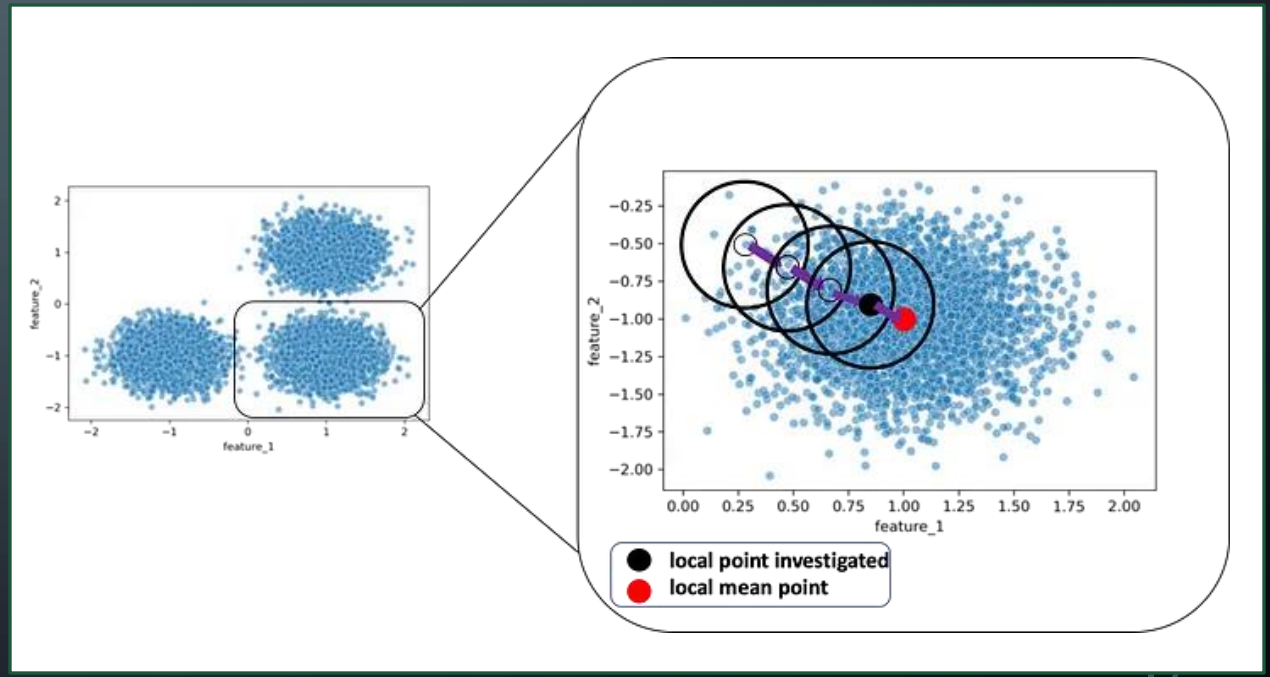
- Napoved bo smiselna le na podatkih, ki so podobni vhodnim
- Pogosto problem z interpretacijo – model ne pojasnjuje, na kakšen način je klasificiral
- Dodaten plus, če zna model povedati s kakšno verjetnostjo poda napoved

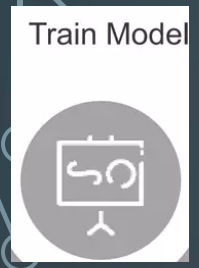


CLUSTERING



- Mean-shift gručenje
  - Ne definiramo števila razdredov
- Težave z interpretacijo delitve
- Kdaj končati algoritem? Ni trivialno





TRENIRANJE MODELA

- Stochastic gradient descent
  - Spreminjanje parametrov modela glede na lokalni gradient (v hiperprostoru parametrov)
  - Stohastičnost - ne pregledamo celotnega prostora, pač pa začnemo v naključni točki.