

K-najbližih suseda

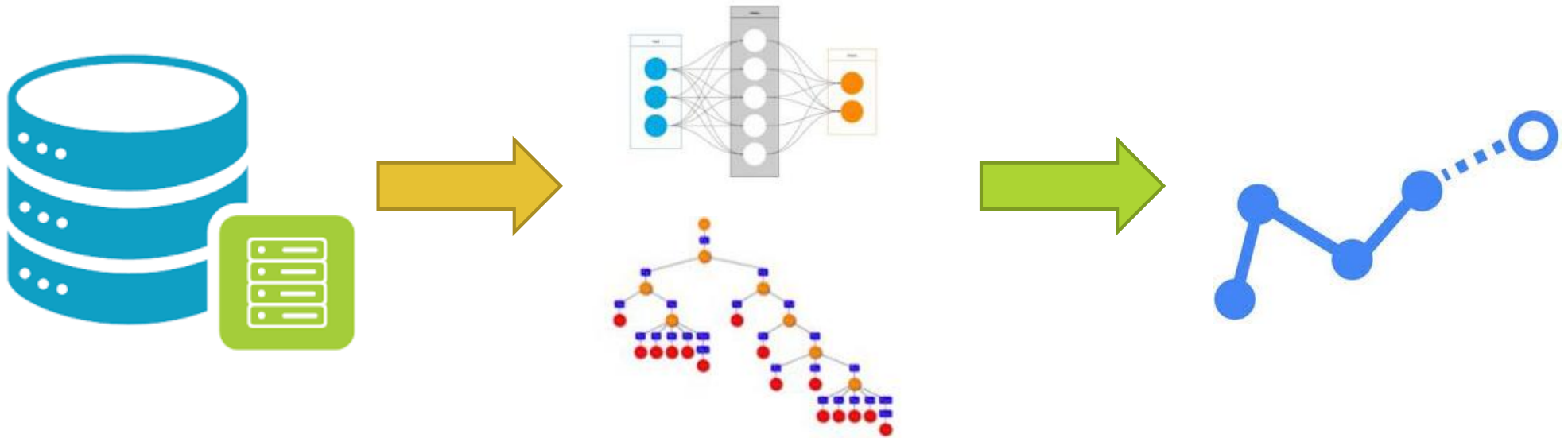
K-NEAREST NEIGHBORS

K-najbližih suseda

- ▶ Šta predstavlja kNN i zašto nam je potreban?
- ▶ Kako najbolje odabrati faktor K?
- ▶ Kada ćemo koristiti kNN algoritam?
- ▶ Kako izgleda i radi kNN algoritam?
- ▶ Analiza podataka korišćenjem kNN

Modeli mašinskog učenja (1)

- ▶ Izračunati prediktivnu vrednost na osnovu učenja iz postojećih podataka, koji su nam dostupni

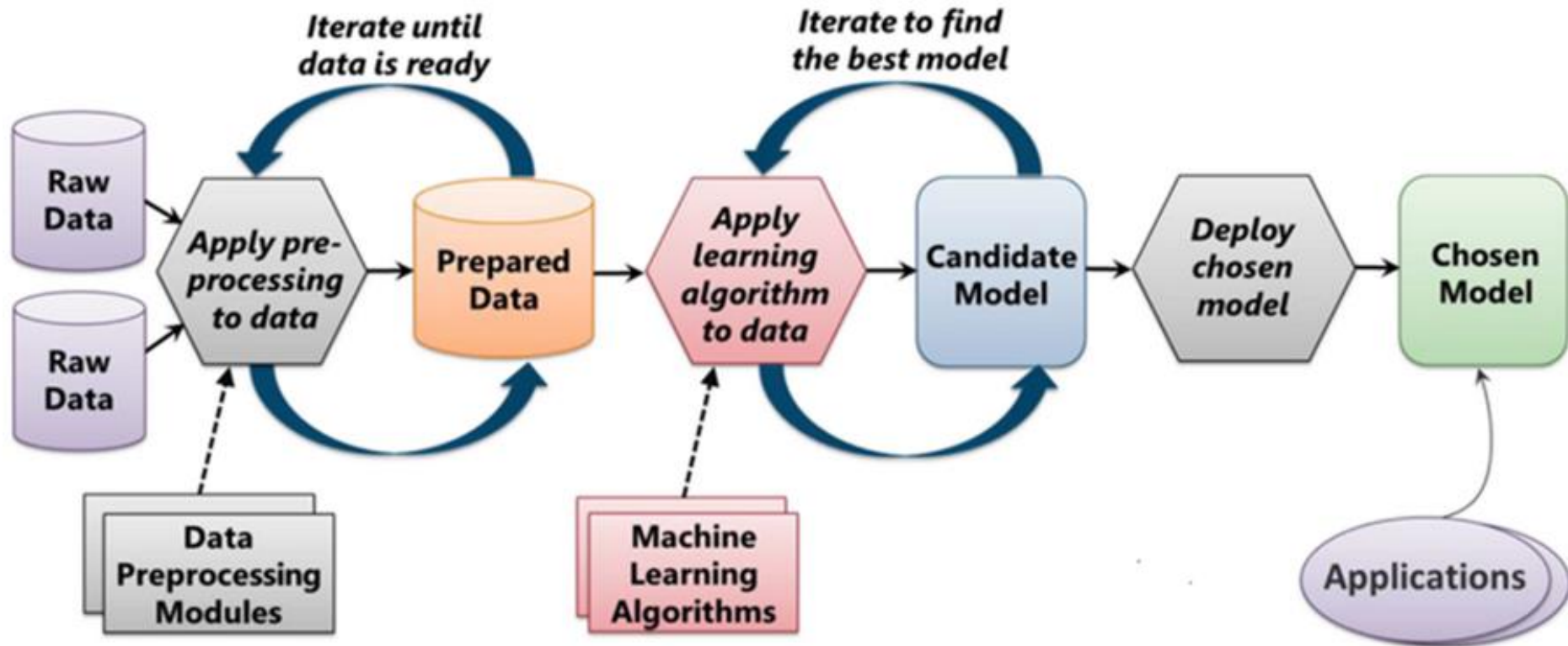



Podaci (ulaz)

Primena neke tehnike ili tehnika
mašinskog učenja

Predviđeni izlaz

Modeli mašinskog učenja (2)





Is that a dog?

No dear, you can
differentiate
between a cat
and a dog based
on their
characteristics

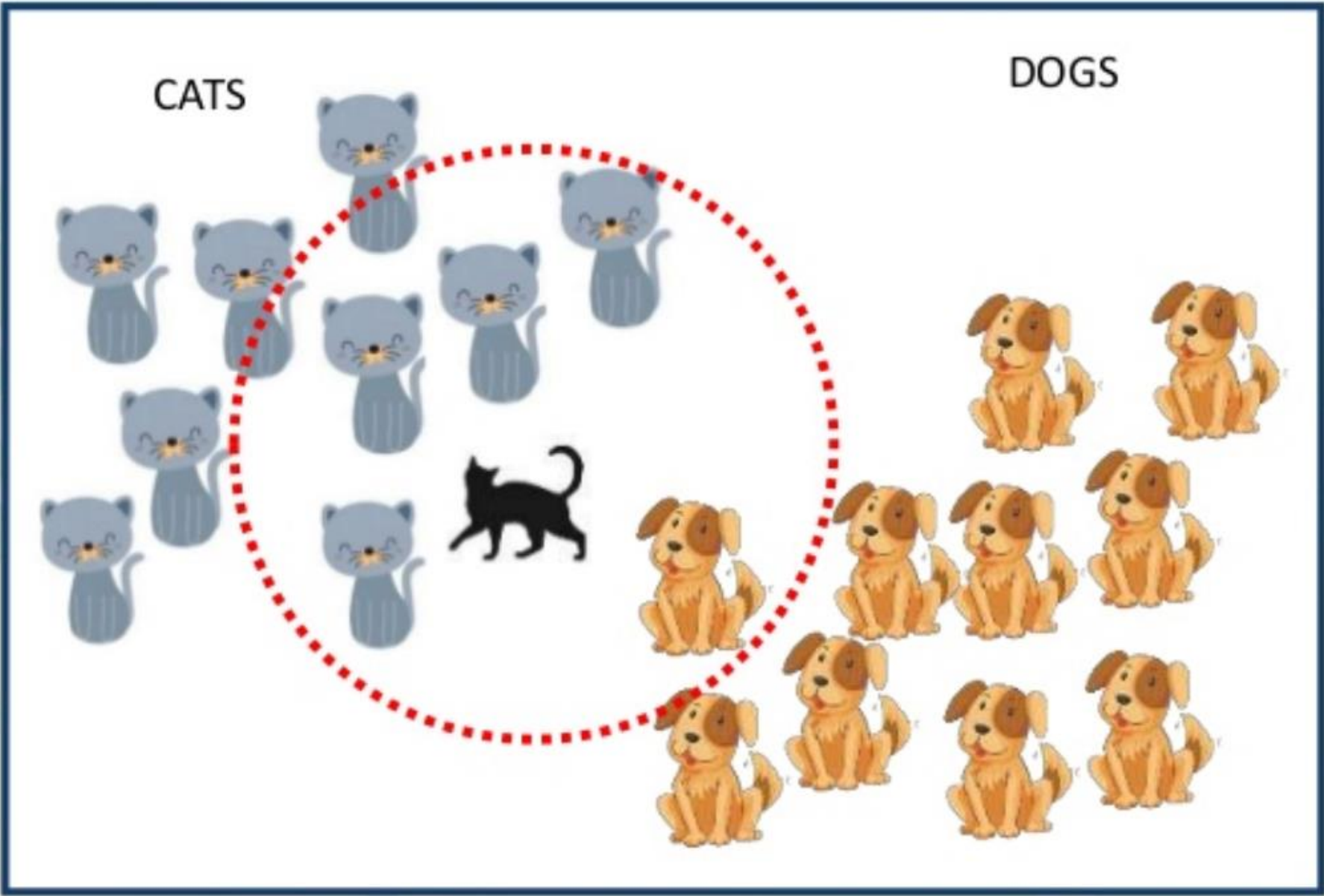
Primer 1

Mačke	Psi
Oštre kandže na šapama, pogodne za penjanje	Meke šape, ne može da se penje uz drvo
Vrlo male uši	Velike uši
Majukanje (MJAU-MJAU)	Lajanje i režanje (AV-AV)
Ne voli da se igra, već da se mazi	Voli da trči i da se igra u društvu
Voli da jede ribu	Voli da jede meso i koske, ne jede ribu

CATS

DOGS

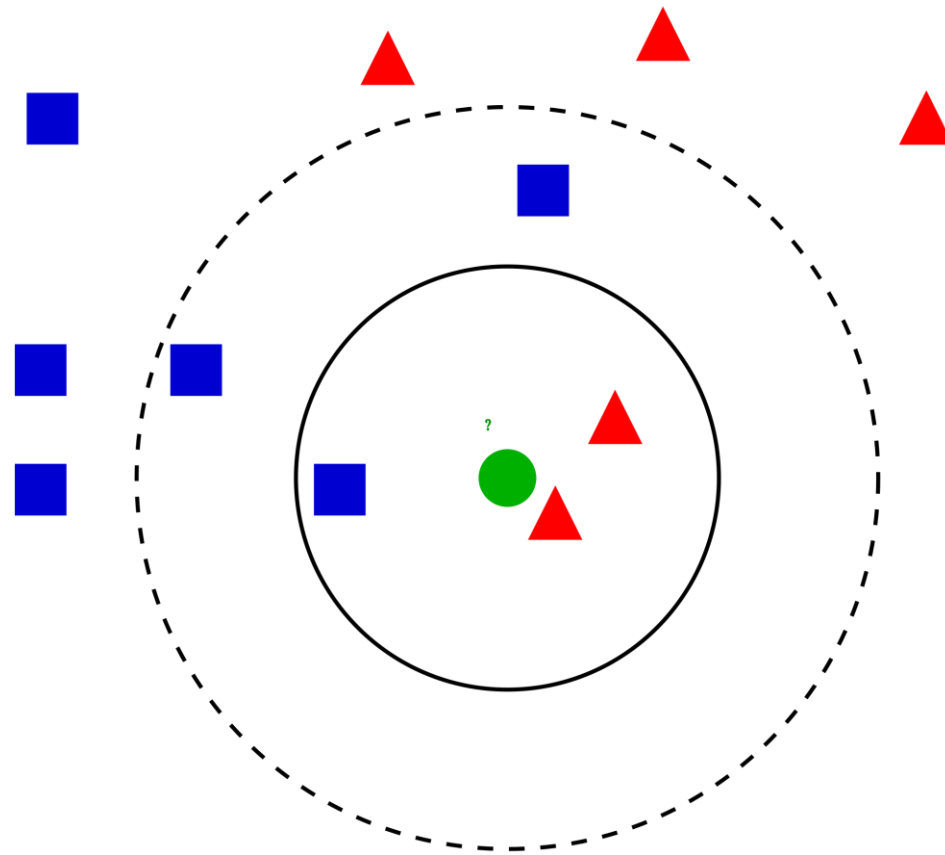
Sharp of claws →



Length of ears →

Primer 2 - Oglasi

- ▶ U svetu digitalnog marketinga:
kako odabrati da li kupcu nuditi artikal X ili Y ?
- ▶ Jedan od načina koristiti klasifikacione tehnike
- ▶ Uticaj ulaznih parametara na izlaz/predikciju

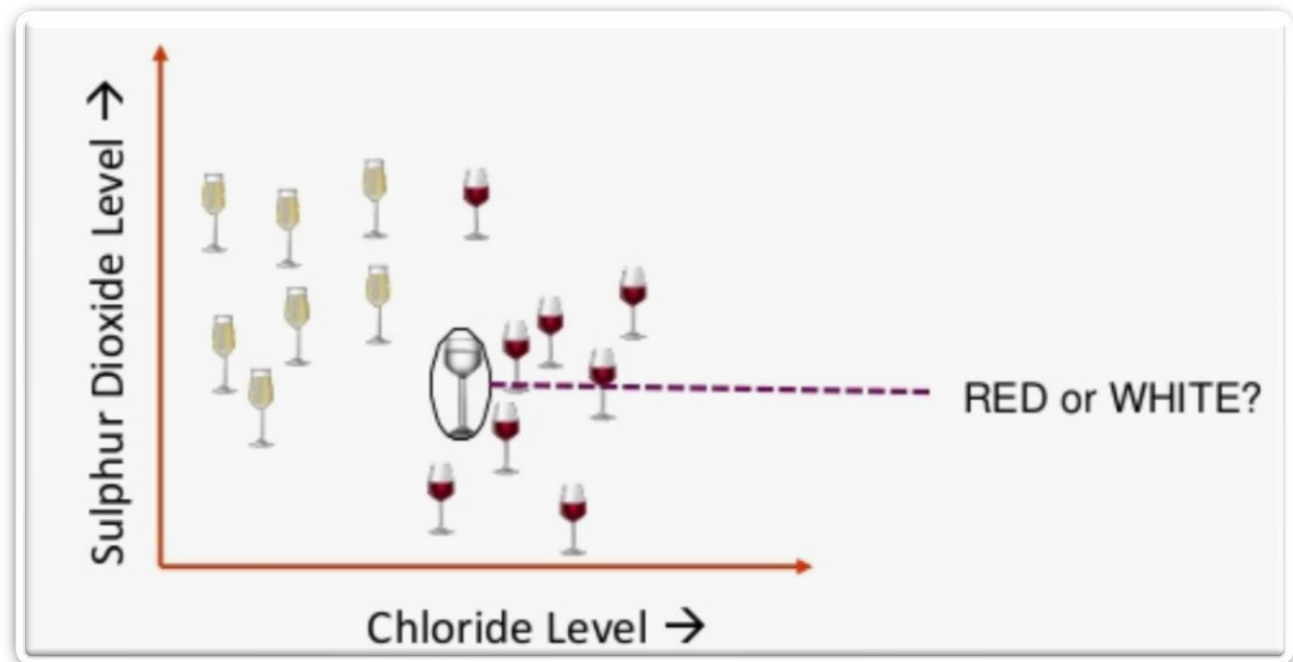


Šta je kNN?

- ▶ Algoritam mašinskog učenja, spada u tehnike nadgledanog učenja (eng. *supervised*)
- ▶ Može se koristiti u klasifikaciji i regresiji, ali su u industriji češći klasifikacioni problemi predikcije
- ▶ Svaki algoritam mašinskog učenja treba da vodi računa o 3 aspekta:
 - ▶ Jednostavno tumačenje izlaznih podataka
 - ▶ Vreme izvršavanja algoritma
 - ▶ Koliko je dobra moć predviđanja
- ▶ kNN u klasifikaciji: klasifikuje tačku posmatranja u odnosu na to kako su susedi klasifikovani

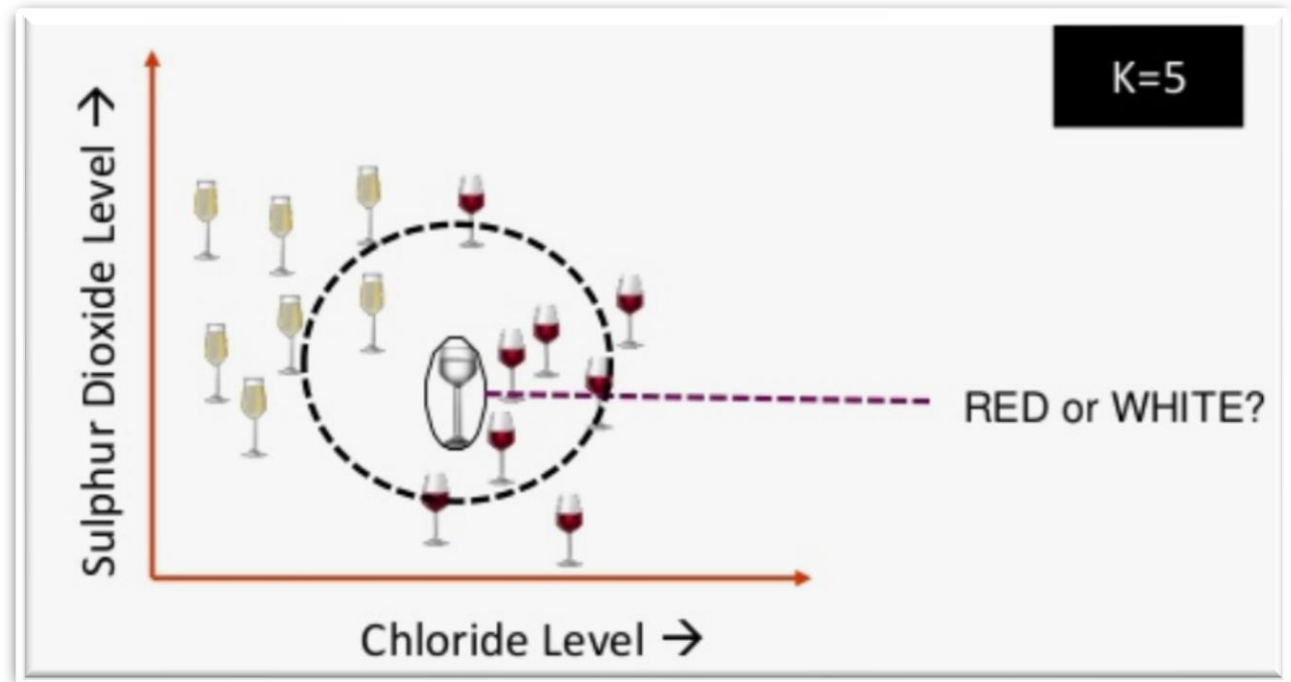
Primer 3 - Belo ili crveno vino?

- ▶ kNN posmatra sve primerke u skupu podataka (*dataset-u*) i klasifikuje nove primerke zasnovane na sličnosti
- ▶ Kolika je vrednost K?



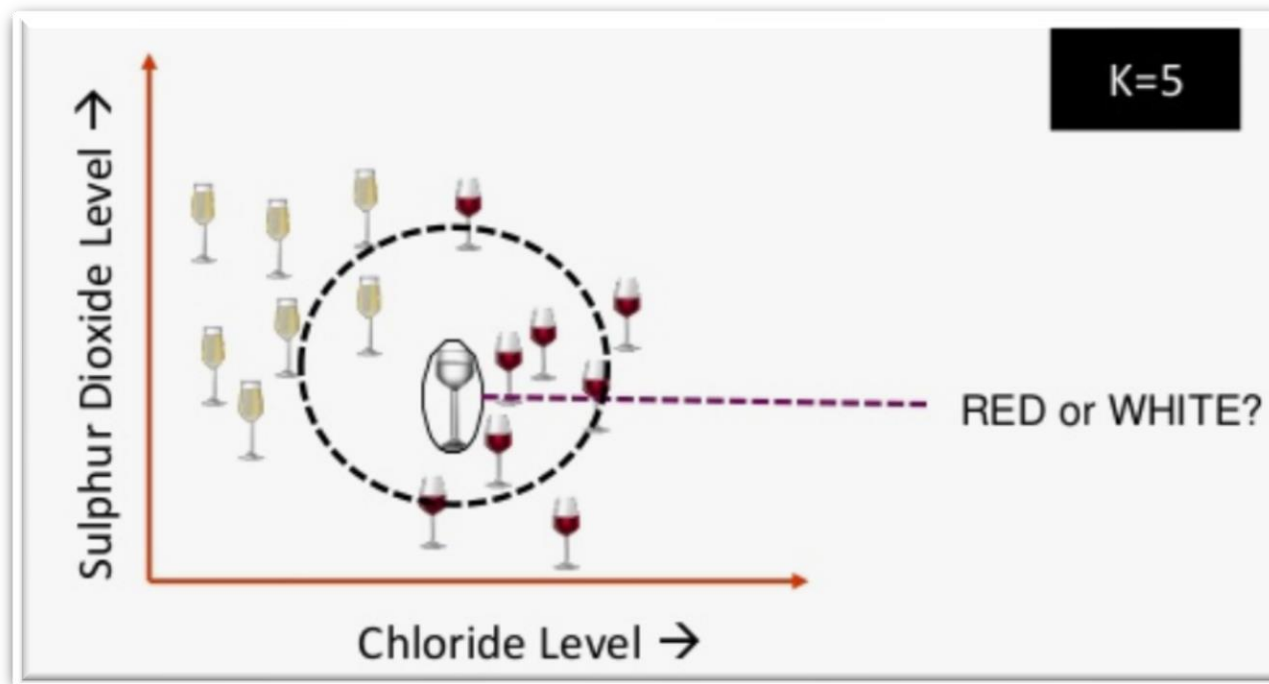
Primer 3 - Belo ili crveno vino?

- ▶ k u KNN algoritmu predstavlja parametar koji označava broj najbližih suseda koji su uključeni (u procesu glasanja)
- ▶ U datom primeru odabraćemo na primer 5 najbližih suseda i na osnovu njih videti koliko primeraka belog ili crvenog vina imamo u tom skupu
- ▶ Vino ćemo klasifikovati kao **crveno**, jer 4 od 5 najbližih suseda pripada kategoriji crvenih vina, a samo jedan susjed pripada belim vinima

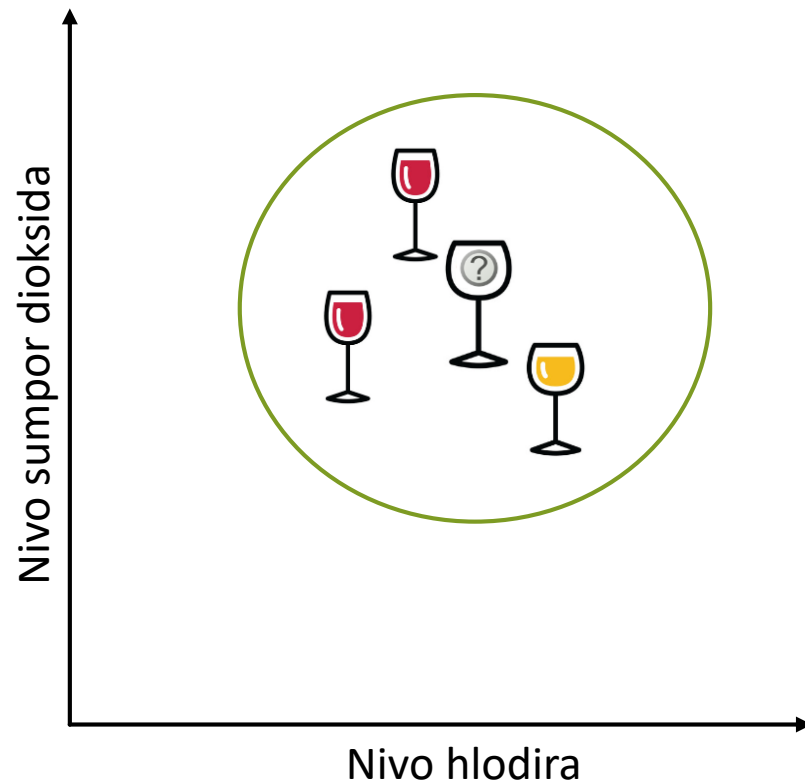


Kako odabrati faktor K?

- ▶ kNN algoritam je zasnovan na sličnosti karakteristika (osobina)
- ▶ Izbor prave vrednosti za faktor K je proces koji se naziva "**podešavanje parametara**" (eng. *parameter tuning*)
- ▶ Što bolji faktor K odredimo, preciznost našeg modela će biti bolja



Primer 3 - Izbor faktora K? (1)

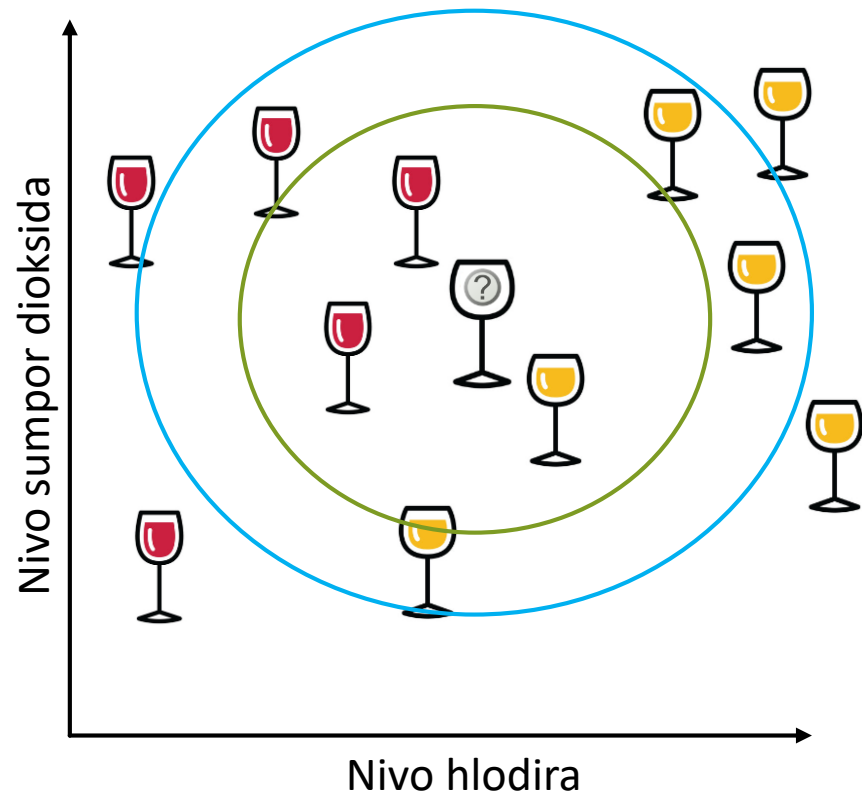


Ukoliko je faktor $k = 3$,
možemo novo vino
u našem skupu podataka
klasifikovati kao:



Crveno: 2
Belo: 1

Primer 3 - Izbor faktora K? (2)



Ukoliko je faktor $k = 7$,
možemo novo vino
u našem skupu podataka
klasifikovati kao:



Crveno: 3
Belo: 4

Primer 3 - Kako odabrati faktor K?

- ▶ Ako je faktor $K=3$ naše vino je crveno, a ako K promenimo na vrednost 7 suseda, onda je belo, koje K odabrati?
- ▶ Izbor faktora K :
 - ▶ \sqrt{n} , gde je n ukupan broj svih instanci (primeraka) u skupu podataka
 - ▶ bira se neparna vrednost k , da bi se izbegla konfuzija kojoj od dve klase pripada
- ▶ **Veća vrednost K ima manju šansu za grešku!**



Kada koristimo kNN algoritam?

► Možemo koristiti algoritam k-najbližih suseda kada:

► imamo podatke koji su označeni



Crveno vino

► imamo podatke "bez šuma"

► imamo mali skup podataka koje analiziramo (zato što je kNN "lazy learning" metoda)

Težina	Visina	Klasa za kilažu osobe
51	167	mršava
62	182	odbojkašica
69	176	23 god.
64	173	normalna
65	172	normalna

Primer 4 - Visine i težine osoba (1)

- ▶ Imamo skup podataka sa visinama (u cm) i težinama (u kg)
- ▶ Izlazne klase su: normalna ili mršava osoba

Težina [kg]	Visina [cm]	Klasa
51	167	mršava
62	182	normalna
69	176	normalna
64	173	normalna
65	172	normalna
56	174	mršava
58	169	normalna
57	173	normalna
55	170	normalna

Primer 4 - Visine i težine osoba (2)

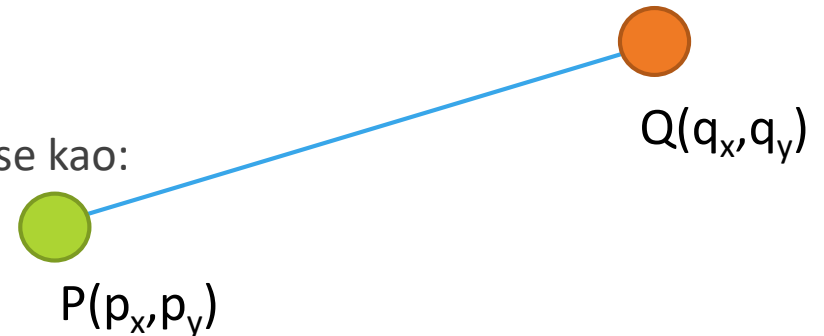
- ▶ Na osnovu datog skupa podataka, a korišćenjem kNN algoritma, odrediti da li ove osobe spadaju u klasu Normalna ili Mršava
- ▶ Pretpostavka: ne znamo kako se računa BMI (*body mass index*) !!!

Osoba	Težina [kg]	Visina [cm]	Klasa
A	57	170	?
B	66	182	?

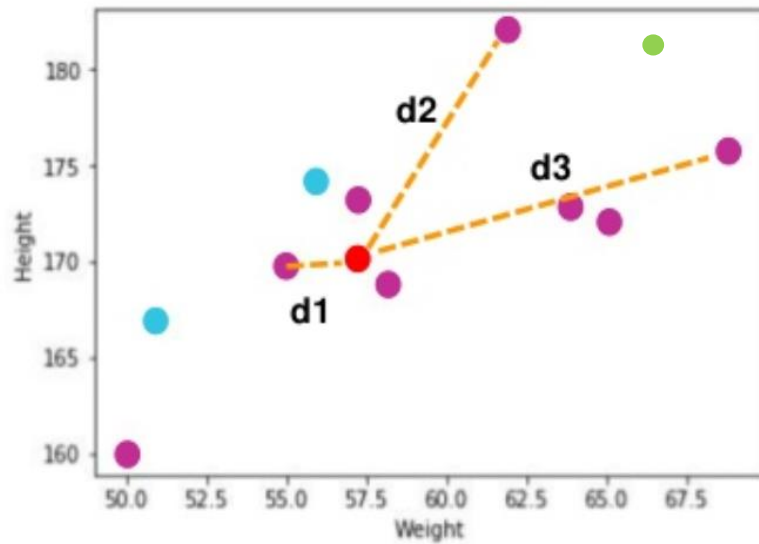
kNN - A ko su nam najbliži susedi?

- ▶ Za pronalaženje najbližih suseda koristimo neku metriku:
 - ▶ Euklidska razdaljina / Pitagorino rastojanje (eng. *Euclidean distance*)
 - ▶ Menhetn razdaljina (eng. *Manhattan distance*)
 - ▶ Čebiševa razdaljina (eng. *Chebyshev distance*)
 - ▶ Mahalanobisova razdaljina (eng. *Mahalanobis distance*)
- ▶ Euklidsko rastojanje u dvodimenzionalnom prostoru računa se kao:

$$dist = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$



Primer 4 - Visine i težine osoba (3)



$$\text{dist}(\mathbf{d1}) = \sqrt{(170-167)^2 + (57-51)^2} \approx 6.7$$

$$\text{dist}(\mathbf{d2}) = \sqrt{(170-182)^2 + (57-62)^2} \approx 13$$

$$\text{dist}(\mathbf{d3}) = \sqrt{(170-176)^2 + (57-69)^2} \approx 13.4$$

● Nepoznata tačka - osoba A

● Nepoznata tačka - osoba B

Primer 4 - Visine i težine osoba (4)

- ▶ Na osnovu tačke $A(x,y) = (57, 170)$ tražimo Euklidsku razdaljinu do svih drugih tačaka, pa na osnovu faktora K određujemo koji susedi su najbliži

Težina [kg]	Visina [cm]	Klasa	Euklidska raz. (od tačke A)
51	167	mršava	6.7
62	182	normalna	13
69	176	normalna	13.4
64	173	normalna	7.6
65	172	normalna	8.2
56	174	mršava	4.1
58	169	normalna	1.4
57	173	normalna	3
55	170	normalna	2

Imamo $n=10$ instanci,
 $\sqrt{10} = 3.1$

Prema faktoru $K=3$,
ovo su 3 najbliža suseda
=> **normalna**

Pseudo kod algoritma kNN

- ▶ 1. Učitaj podatke
- ▶ 2. Izračunaj vrednost faktora K
- ▶ 3. Da dobijete prediktivnu klasu, iterirajte od prvog do poslednjeg primerka u okviru trening skupa podataka
 - ▶ 3.1. Izračunati razdaljinu između tražene instance (tačke) i svih instanci iz trening skupa. (u našem primeru smo odabrali Euklidsko rastojanje)
 - ▶ 3.2. Sortirati izračunate razdaljine u rastućem poretku.
 - ▶ 3.3. Uzeti najviših K redova iz sortiranog niza.
 - ▶ 3.4. U K redova analizirati koliki je broj pojavljivanja klasa.
 - ▶ 3.5. Dobijeni rezultat je prediktivna klasa za traženu instancu.

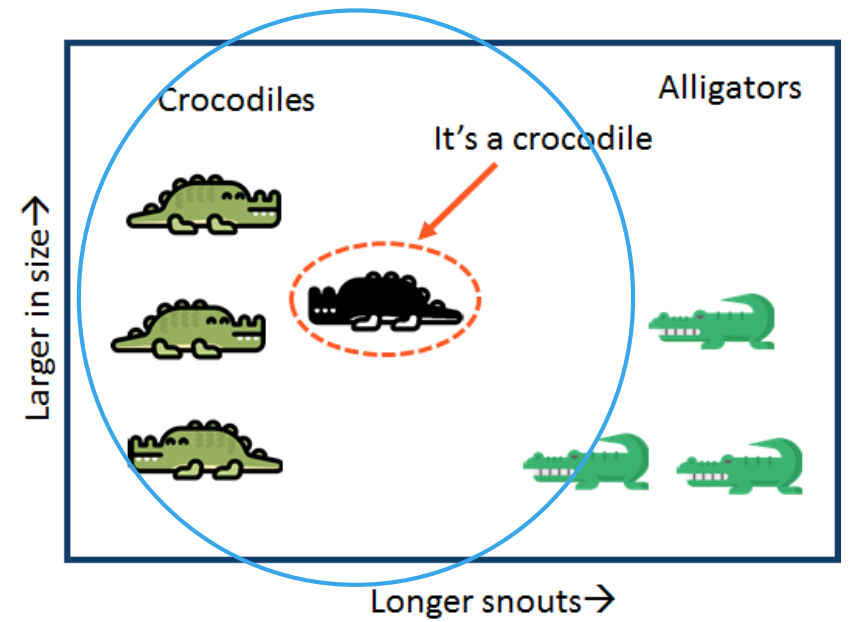
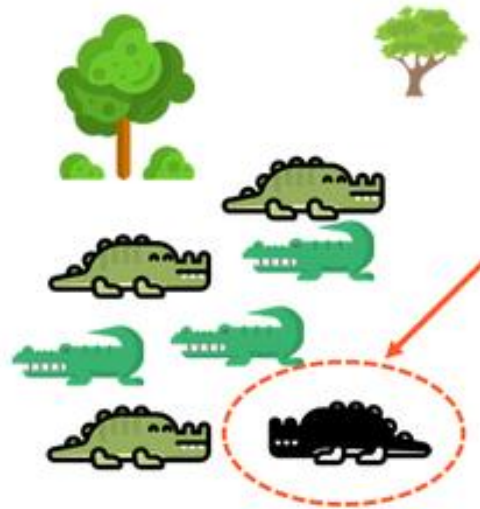
Česta greška: kNN i k-Means poistovećivanje

	kNN (k najbližih suseda)	k-Means (k srednjih vrednosti)
Tehnika učenja	nadgledano (<i>supervised</i>)	nenadgledano (<i>unsupervised</i>)
Za koje probleme se koristi	klasifikaciju (i ponekad regresiju)	klasterizaciju
Na čemu je zasnovan	sličnosti karakteristika	deljenje objekata u klasteru (tako da svaka instanca mora pripadati tačno jednom klasteru)
Šta znači K	k je broj najbližih suseda korišćenih u klasifikaciji	k je broj klastera koje algoritam pokušava da identifikuje iz podataka
Gde se koristi	koristi se za klasifikaciju i regresiju poznatih podataka, gde je obično atribut/vrednost uslova ciljane instance poznata (samo nije prediktivna klasa poznata)	koristi se obično za scenarije razumevanja demografske slike stanovništva, segmentacije tržišta, trendova u društvenim medijima, otkrivanja anomalija,... (svuda gde su klasteri nepoznati)

Česta greška: kNN i k-Means poistovećivanje (2)

	kNN (k najbližih suseda)	k-Means (k srednjih vrednosti)
Treniranje	<p>kNN nema klasičnu fazu obučavanja, a predviđanje se vrši na osnovu samo K najbližih suseda (često sa Euklidskim rastojanjem) na osnovu ponderisanih vrednosti tačaka koje se posmatraju.</p> <p>Algoritam završava rad kada se sve instance koje se posmatraju klasifikuju (sa željenom tačnošću).</p>	<p>k-Means u fazi obučavanja posmatra K odabranih instanci (~ centroidi). Svaka tačka u vektorskom prostoru je dodeljene klasteru koji predstavlja najbliže (euklidsko rastojanje) od centroida.</p> <p>Kada se klasteri formiraju, za svaki klaster centroid se ažurira na srednju vrednost svih članova klastera. Formiranje klastera ponovo počinje (resetuje se) sa novim centroidom. Ovo se ponavlja sve dok centroidi sami ne postanu klasteri.</p> <p>Predviđanje se vrši na osnovu najbližeg centroida.</p>
Podaci	kNN zahteva označene tačke (labelirane)	k-Means ne zahteva označene tačke

kNN



k-Means

