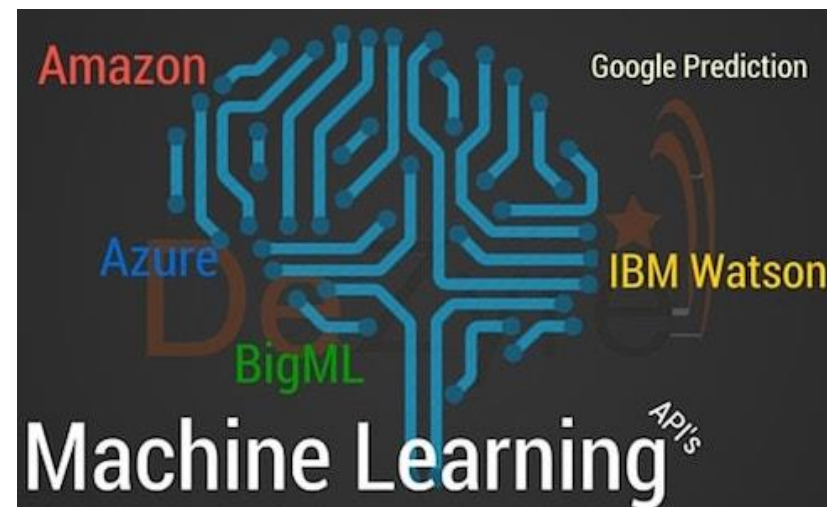
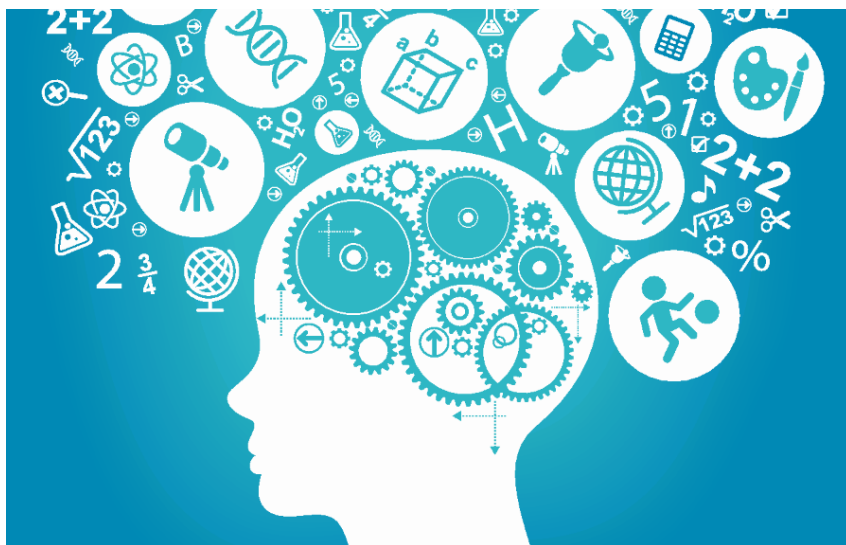




STROJNO UČENJE



ŠTO JE STROJNO UČENJE?

- Strojno učenje podučava računalne sustave ono što je ljudima prirodno:

Učenje iz iskustva

- Algoritmi strojnog učenja primjenjuje računalne metode koje “uče” informacije direktno iz podataka, bez oslonca na teorijske jednadžbe i modele;
- Algoritmi se **poboljšavaju** kako raste broj uzoraka dostupnih za učenje.

VIŠE PODATAKA, VIŠE PITANJA, BOLJI ODGOVORI

- Algoritmi strojnog učenja pronalaze **prirodne uzorke** u podacima te na temelju toga dobivaju uvid i zatim **odlučuju i predviđaju**;
- Primjenjuju se već svakodnevno za donošenje ključnih odluka u **medicinskoj dijagnostici, trgovanju dionicama, predviđanju potrošnje energije**, itd.
- Web stranice mnogih medija oslanjaju se na strojno učenje kako bi vam iz milijuna mogućnosti **“prosijale”** i preporučile npr. pjesmu ili film;
- Trgovci ih koriste kako bi stekli uvid u navike svojih kupaca.

PRIMJENA

Porastom količine podataka koji se kontinuirano stvaraju strojno učenje postalo je posebno važno:

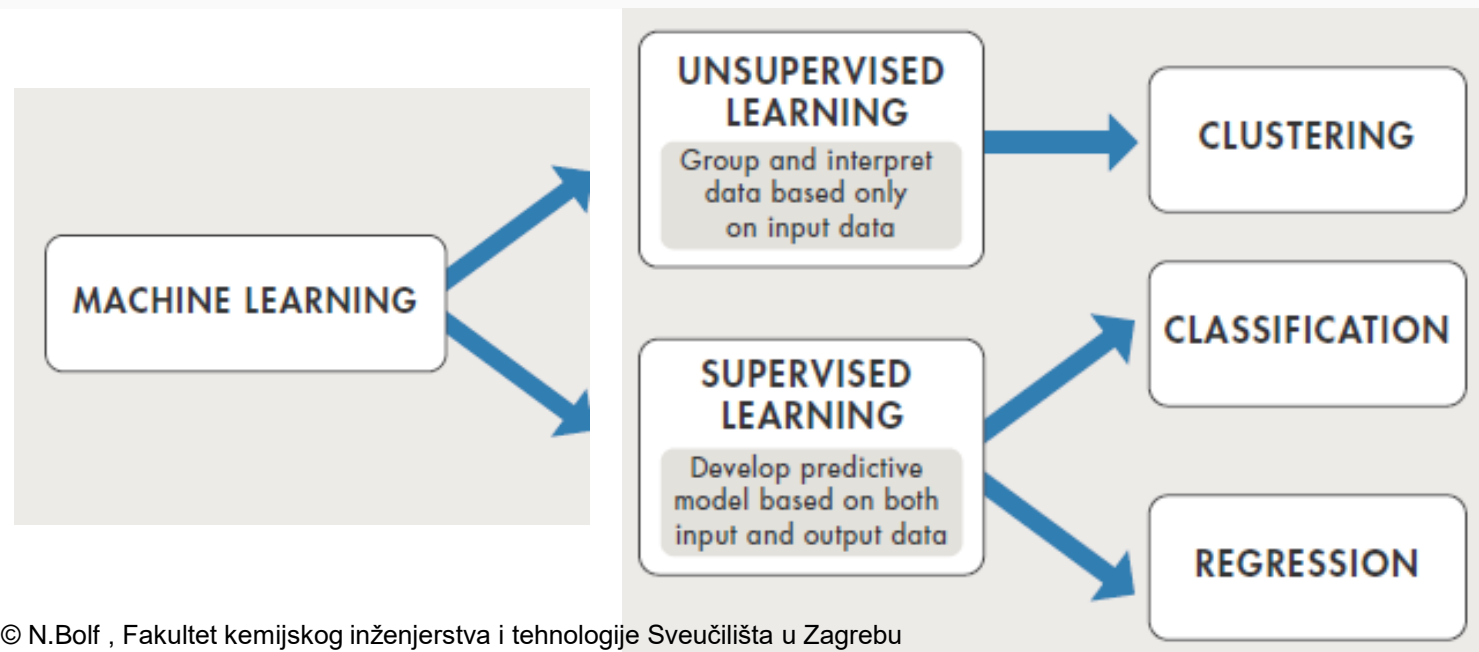
- **Računalne financije** – procjena kredita i algoritamsko trgovanje (burza)
- **Obrada slika i računalna vizija** – prepoznavanje lica, detekcija gibanja i detekcija objekata
- **Računalna biologija** – detekcija tumora, istraživanje lijekova i sekvenciranje DNA
- **Proizvodnja energije** – prognoza cijena i opterećenja mreže
- **Industrijska proizvodnja**
prediktivno održavanje
- **Obrada prirodnog jezika**



KAKO FUNKCIONIRA STROJNO UČENJE

Osnovne tehnike:

- **Nadzirano učenje (Supervised learning)**
Uvježbavanje modela s poznatim skupom ulaza i izlaza tako da mogu predviđati buduće izlaze na temelju dostupnih ulaza;
- **Nenadzirano učenje (Unsupervised learning)**,
Pronalazi skrivene uzorke i inherentne strukture u podacima.



NADZIRANO UČENJE

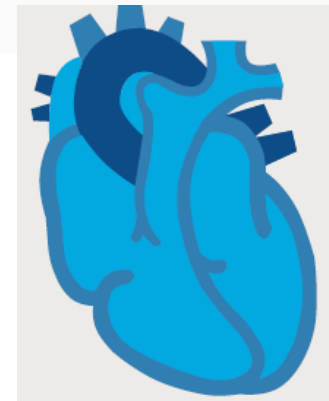
Algoritmi uzimaju poznati skup **ulaznih podataka** i **poznate odzive (izlaze)** i uvježbavaju model za predviđanje (**prediction**).

Primjena

- **Klasifikacija** - predviđanje **diskretnih odziva**
 - Ulazni podaci klasificiraju se u **kategorije**;
 - npr. je li mail originalni ili neželjena pošta (spam),
medicinska dijagnostika - je li tumor kancerogen ili dobroćudan,
prepoznavanje govora i sl.
- **Regresija** - predviđa **kontinuirane odzive**
 - npr. promjene temperature ili koncentracije u reaktoru
fluktuacija potrošnje električne energije

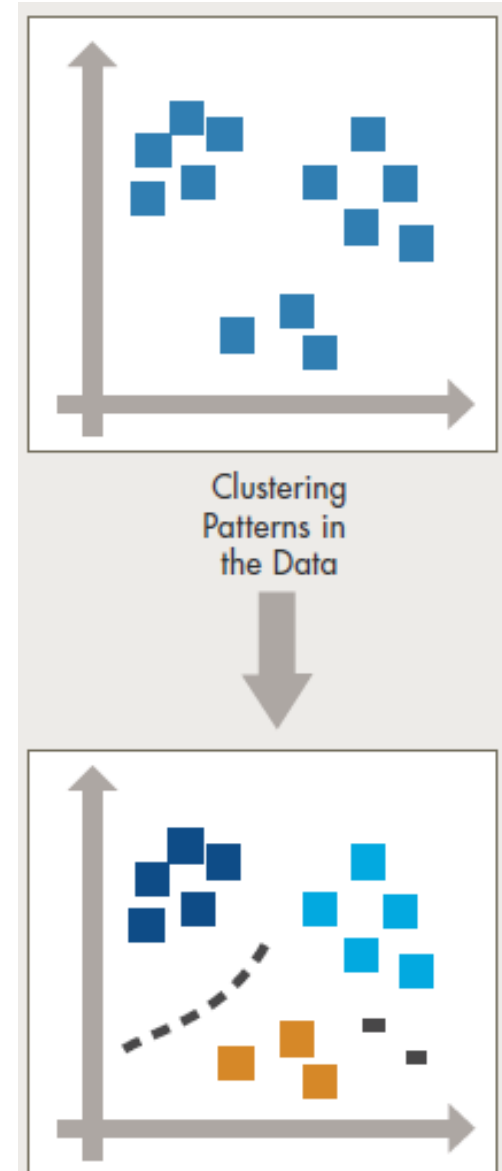
NADZIRANO UČENJE ZA PREDVIĐANJE INFARKTA

- Pretpostavimo da kliničari žele predvidjeti hoće li pacijente doživjeti uskoro **srčani udar**;
- Podaci o pacijentu uključuju **dob, težinu, visinu, krvni tlak** i sl.
- U arhivi postoje informacije o povijesti bolesti prethodnih pacijenata (npr. tipični simptomi prije srčanog udara);
- Zadatak je, prema tome, kombinirati postojeće podatke u model koji može predvidjeti je li novi pacijent u skoro vrijeme u opasnosti od srčanog udara.



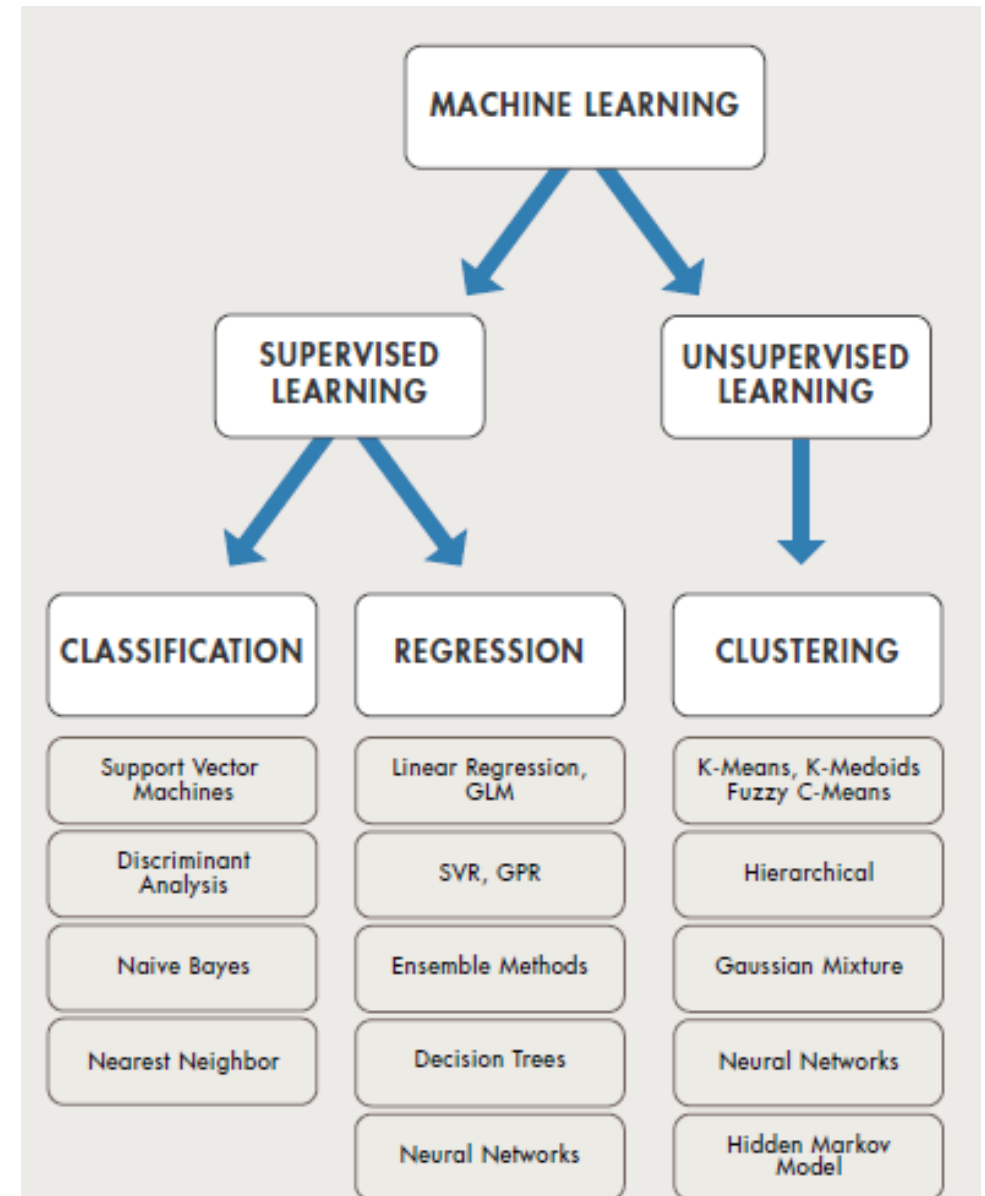
NENADZIRANO UČENJE

- Nenadzirano učenje pronalazi **skrivenne uzorke** ili **intrizičke strukture** u podacima;
- Koristi se za **izvođenje zaključaka** iz skupova podataka koje čine samo ulazni podaci bez odziva (izlaza);
- Najčešća tehnika nenadziranog učenja je **grupiranje (clustering)** pri čemu se traže skriveni obrasci ili grupe;
- Primjena kod **analiza sekvenca gena, analize tržišta, prepoznavanje objekta** itd.



KAKO ODABRATI ALGORITAM?

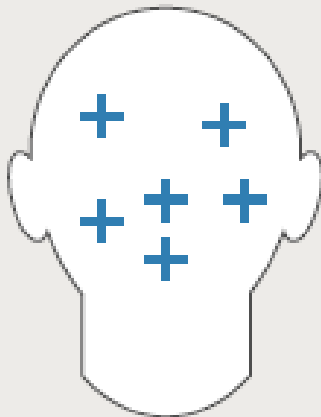
- **Izbor algoritma je velik** – postoji više desetaka algoritama;
- Svaki od njih ima drugačiji pristup učenju;
- Ne postoji najbolja niti univerzalna metoda;
- Čak i vrloiskusni eksperti ne mogu predvidjeti hoće li algoritam raditi dobro (**pokušaj i pogreška!**);
- Odabir algoritma ovisi i o **broju** i **vrsti** podataka, što se želi postići te gdje i kako će se **primijeniti** rezultati.



KADA PRIMIJENITI STROJNO UČENJE?

- Strojno učenje treba primijeniti kod rješavanja **kompleksnih zadataka** ili problema koji uključuju **veliku količinu podataka** i **puno varijabli**, ali ne postoje formule ili jednadžbe (modeli).

Hand-written rules and equations are too complex—as in face recognition and speech recognition.



The rules of a task are constantly changing—as in fraud detection from transaction records.



The nature of the data keeps changing, and the program needs to adapt—as in automated trading, energy demand forecasting, and predicting shopping trends.



PRIMJER ANALIZE UMJETNIČKIH DJELA

Analiza umjetnička djela

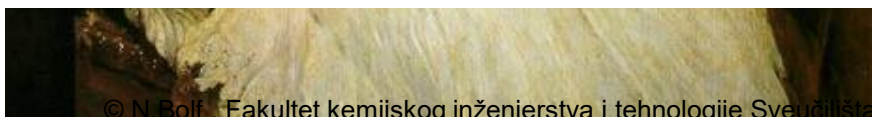
- Istraživači *Laboratorija za umjetnost i umjetnu inteligenciju* na *Sveučilištu Rutgers* željeli su ustanoviti može li računalni algoritam **klasificirati** slike prema **stilu, žanru i umjetniku**;
- Započeli su prepoznavanjem **vizualnih značajki** za klasificiranje stila slikara;
- Algoritmi koje su razvili klasificiraju **stilove slika** u bazi podataka s **60% točnosti**, što nadmašuje tipične laike;
- Istraživači su pretpostavili da se vizualne značajke korisne za **klasifikaciju stila (nadzirano učenje)** također mogu koristiti za određivanje umjetničkih utjecaja (**nenadzirano učenje**);
- Primijenili su **algoritme za klasifikaciju** koji su učili na slikama s *Google-a*;
- Ispitali su algoritme na više od **1.700** slika od **66** različitih umjetnika koji su stvarali u rasponu od **550** godina.

VELAZQUEZ VS. BACON

- Algoritam je lako prepoznao povezana djela, npr. utjecaj *Portreta pape Inocenta X* **Diega Velazqueza** na *Studiju Velazquezovog portreta pape Inocenta X* **Francisa Bacona**.



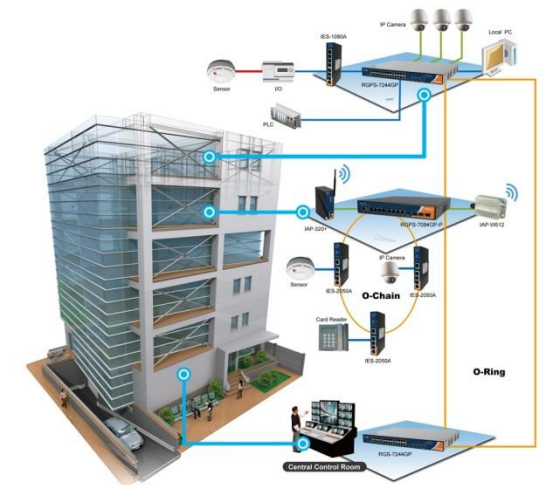
Ključan korak u procesu klasifikacije jest preslikavanje informacija o karakteristikama slikarskog djela kao što su kombinacija boja, modulacija svjetlosti i sjene, mekoća rubova, jedinstveni slikarski rukopis sadržan u potezima kista, materičnosti namaza boje i karakterima linija u skup numeričkih veličina dobivenih primjenom raznih metoda za ekstrakciju značajki slike.



PRIMJER UŠTEDE ENERGIJE

Optimiranje potrošnje energije u zgradama

- Sustavi za grijanje, ventilaciju i hlađenje (**HVAC**) u poslovnim zgradama, bolnicama i velikim trgovačkim centrima često su nedjelotvorni jer ne uzimaju u obzir promjene vremena, varijabilne troškove električne energije i toplinska svojstva zgrada;
- **Inteligenta “cloud” platforma** rješava taj problem;
- Platforma primjenjuje **napredne algoritme** i **strojno učenje** za kontinuiranu obradu informacija s električnih brojila, mjerila temperature i tlaka, isto kao i vremenske prognoze i cijene energije;
- Strojno učenje **segmentira podatke** i određuje **relativni doprinos** potrošnje plina, struje, pare i solarne energije na proces grijanja i hlađenja;
- Tijekom normalnog rada ovi sustavi smanjuju potrošnju energije HVAC sustava u velikim objektima od **10% do 25%**.



ZA ONE KOJI ŽELE VIŠE...

POGLEDAJTE u Matlabu

Machine Learning Made Easy (34:34)

Signal Processing and Machine Learning Techniques for Sensor Data Analytics (42:45)



ČITAJTE

Machine Learning Blog Posts: Social Network Analysis, Text Mining, Bayesian Reasoning, ...

The Netflix Prize and Production Machine Learning Systems: An Insider Look

Machine Learning Challenges: Choosing the Best Model and Avoiding Overfitting

ISTRAŽITE

MATLAB Machine Learning Examples

Machine Learning Solutions

Classify Data with the Classification Learner App

NETFLIXOVA NAGRADA



Netflix Prize

From Wikipedia, the free encyclopedia

The **Netflix Prize** was an open competition for the best **collaborative filtering algorithm** to predict user ratings for **films**, based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.

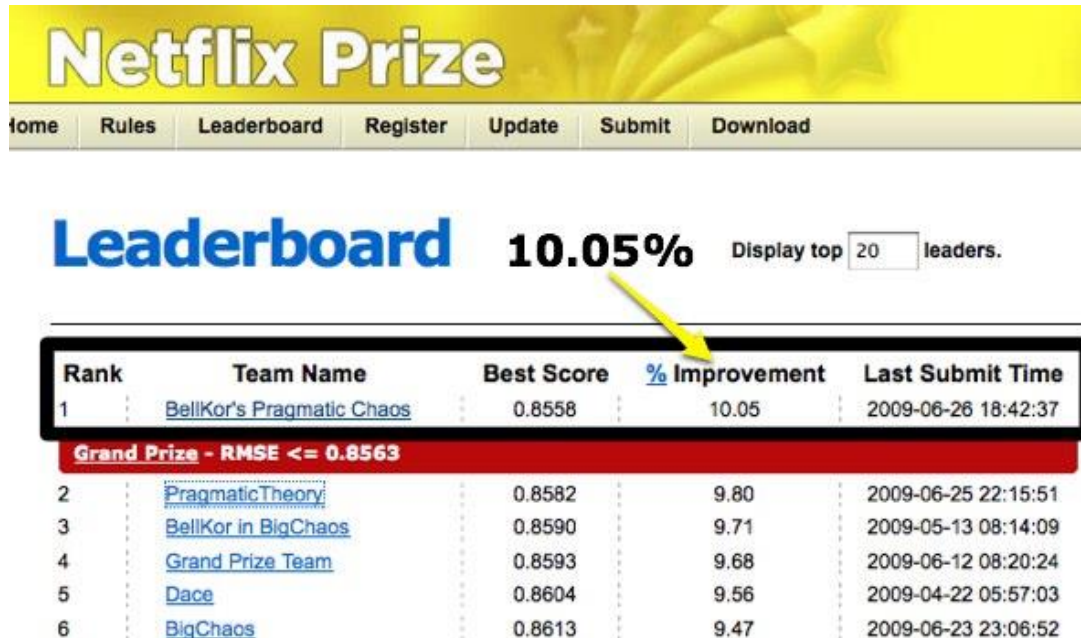
The competition was held by **Netflix**, an online DVD-rental and video streaming service, and was open to anyone who is neither connected with Netflix (current and former employees, agents, close relatives of Netflix employees, etc.) nor a resident of certain blocked countries (such as Cuba or North Korea).^[1] On September 21, 2009, the grand prize of US\$1,000,000 was given to the BellKor's Pragmatic Chaos team which bested Netflix's own algorithm for predicting ratings by 10.06%.^[2]

NETFLIXOVA NAGRADA

- **Netflix** ima na raspolaganju skup podataka kojeg čine **100.480.507** ocjena koje su dali **480.189** korisnika za **17.770** filmova;
- Svaku ocjenu čine 4 komponente:
korisnik, film, datum ocjene, ocjena
- Podaci su formata *integer ID*, a ocjene su od * do *****
- Skup za kvalifikaciju ima **2.817.131** podatak oblika:
korisnik, film, datum ocjene + ocjena poznata samo poroti.
- Algoritam timova koji sudjeluju mora predviđati ocjene na cijelom skupu za kvalifikaciju, no ...
- ... dobivaju informacije samo za pola skupa podataka - **1.408.342**;
- Druga polovica od **1.408.789** za testiranje služi poroti za ocjenu modela i odabir pobjednika.

NETFLIXOVA NAGRADA

- Samo suci znaju koje su ocjene u skupu za provjeru, a koji u skupu za testiranje;
- Predviđanja se ocjenjuju prema stvarnim ocjenama primjenom korijena kvadrata pogreške (**root mean squared error – RMSE**), a cilj je smanjiti pogrešku što je više moguće;
- Iako su stvarne ocjene cijeli brojevi u rasponu od 1 do 5, predviđanja ne moraju biti.



The screenshot shows the Netflix Prize Leaderboard. At the top, there is a yellow banner with the text "Netflix Prize". Below the banner is a navigation bar with links: Home, Rules, Leaderboard, Register, Update, Submit, and Download. The main heading is "Leaderboard" followed by "10.05%" and "Display top 20 leaders." A yellow arrow points to the "10.05%" value. Below this is a table with the following data:

| Rank | Team Name | Best Score | % Improvement | Last Submit Time |
|--|---|------------|---------------|---------------------|
| 1 | BellKor's Pragmatic Chaos | 0.8558 | 10.05 | 2009-06-26 18:42:37 |
| Grand Prize - RMSE <= 0.8563 | | | | |
| 2 | PragmaticTheory | 0.8582 | 9.80 | 2009-06-25 22:15:51 |
| 3 | BellKor in BigChaos | 0.8590 | 9.71 | 2009-05-13 08:14:09 |
| 4 | Grand Prize Team | 0.8593 | 9.68 | 2009-06-12 08:20:24 |
| 5 | Dace | 0.8604 | 9.56 | 2009-04-22 05:57:03 |
| 6 | BigChaos | 0.8613 | 9.47 | 2009-06-23 23:06:52 |

STROJNO UČENJE



- Kod razvoja modela strojnog učenja rijetko sve ide glatko od početka do kraja;
- Obično se iteracijski pokušavaju različiti pristupi i ideje.

IZAZOVI KOD STROJNOG UČENJA

- Većina izazova vezana je uz **baratanje podacima** i pronalazak **pravog modela**;
- Datoteke s podacima javljaju se u puno različitih oblika i veličina;
- Podaci iz stvarnih sustava mogu biti **nestrukturirani, nepotpuni** i u **različitim formatima**;
- Ponekad su to samo jednostavni brojevi. No, često to je kombinacija **nekoliko različitih vrsta podataka**,
(npr. signali sa senzora, tekst, slika s kamera i sl.);
- **Predobrada podataka** obično zahtjeva posebna znanja i alate;
- Npr. da bi odabrali karakteristike algoritma za **detekciju objekta** potrebna su specijalizirana znanja o obradi slike;
- U analizi **trendova procesnih varijabli** potrebna su znanja iz kemijskog inženjerstva i dinamike procesa te analize signala.

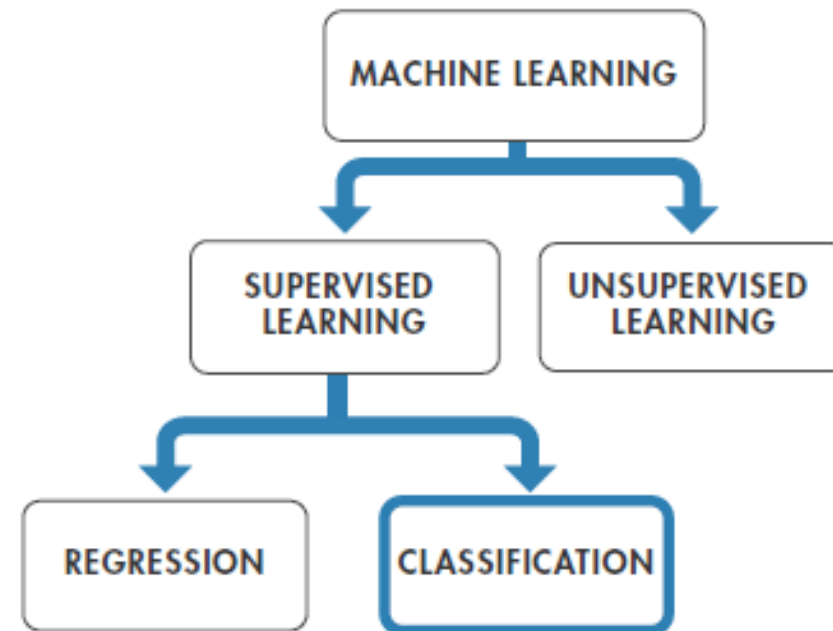
PITANJA NA POČETKU RAZVOJA

Tri ključna pitanja u svakom projektu strojnog učenja:

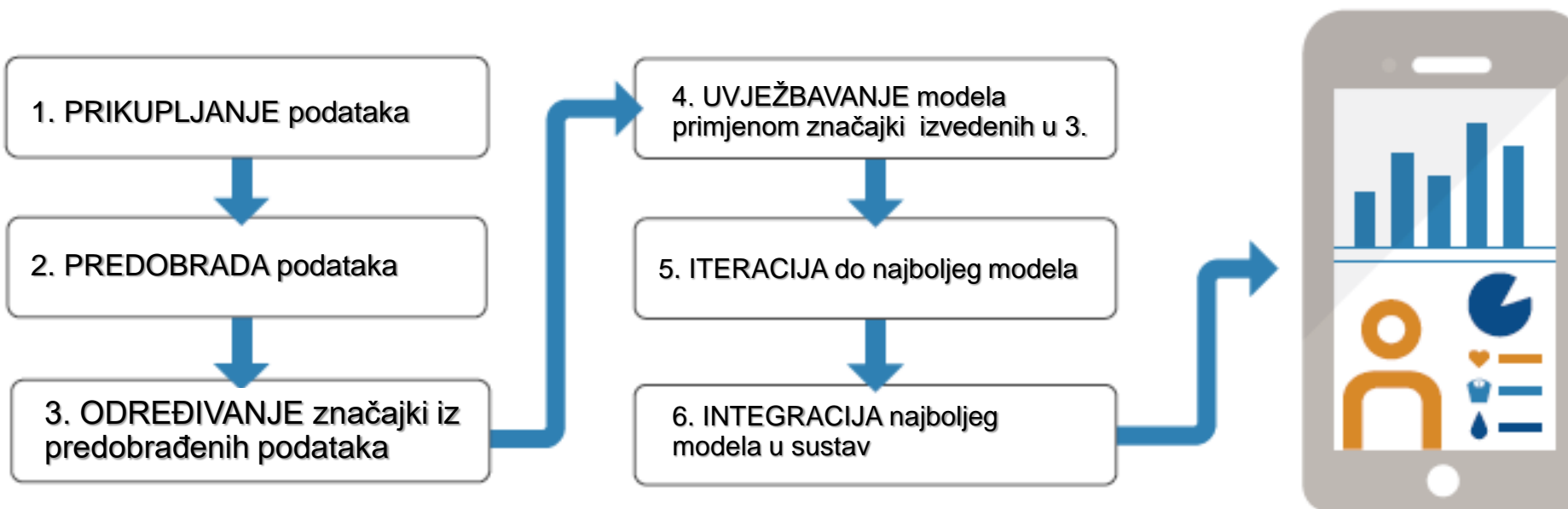
- S kojom **vrstom podataka** radite?
- Koje spoznaje želite **izvući** iz podataka?
- Gdje i kako će se te spoznaje **primijeniti**?

Choose supervised learning if you need to train a model to make a prediction—for example, the future value of a continuous variable, such as temperature or a stock price, or a classification—for example, identify makes of cars from webcam video footage.

Choose unsupervised learning if you need to explore your data and want to train a model to find a good internal representation, such as splitting data up into clusters.

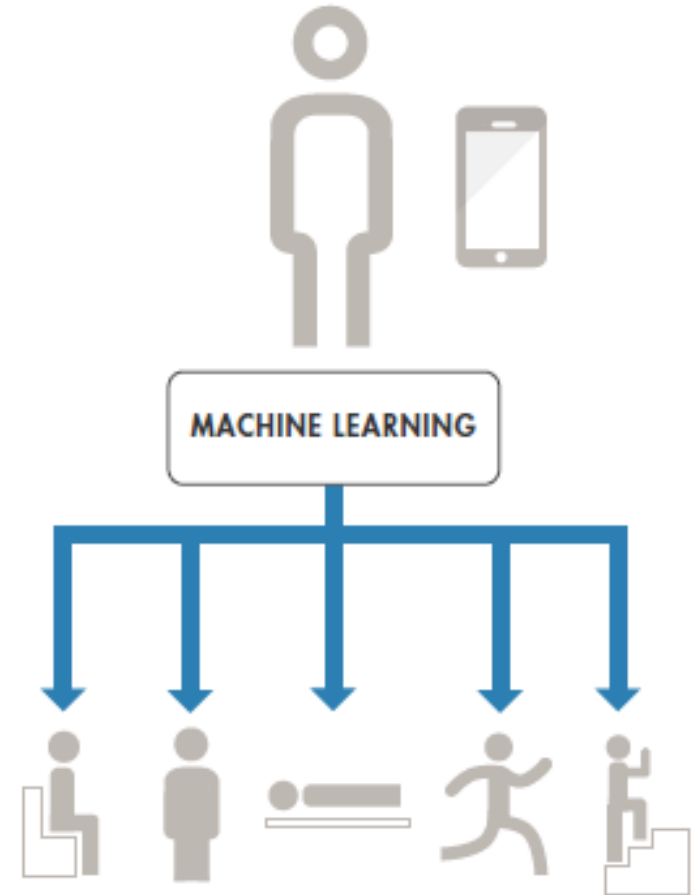


TIJEK RAZVOJA



MODEL ZA PREPOZNAVANJE AKTIVNOSTI

- Ovim primjerom analiziramo aplikaciju za nadgledanje aktivnosti na mobitelu;
- Temelji se na podacima s **3D-senzora akcelerometra i žiroskopa** u mobitelu;
- Odziv (**izlaz**) su aktivnosti: **šetanje, stajanje, trčanje, penjanje po stepenicama, plesanje, ležanje**;
- Ove podatke upotrijebit ćemo za uvježbavanje **klasifikacijskog modela** kako bi identificirali aktivnosti;
- Primijenit ćemo **nadzirano učenje**;
- Izgrađeni model integrira se u aplikaciju kako bi pomogli korisnicima da detektiraju svoje aktivnosti.



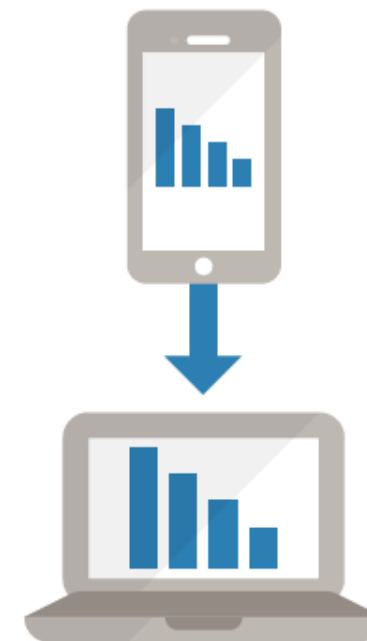
KORAK 1: Učitavanje podataka

Da bi učitali podatke s akcelerometra i žiroskopa radimo sljedeće:

1. Sjednite i držite telefon, učitajte podatke s telefona i spremite ih u datoteku imena “**Sjedenje**”
2. Stojite i držite telefon, zabilježite podatke s telefona, i spremite ih u drugu datoteku “**Stajanje**”
3. Ponovite korake dok ne prikupite podatke za svaku aktivnost koju želite klasificirati;

Podatke s nazivama spremamo u **.txt** datoteku. Format datoteke kao **.txt** ili **.csv** je jednostavan za rad i učitavanje podataka.

- Algoritmi strojnog učenja nisu dovoljno pametni da uoče razliku između šuma i korisne informacije;
- Prije upotrebe podataka za treniranje (učenje), moramo ih **filtrirati** i **kompletirati**.



KORAK 2: Predobrada podataka...

Učitavamo podatke u MATLAB i grafički prikazujemo svaki skup.
Tijekom predobrade podataka radimo slijedeće:

1. **Potražiti *outliere*** – mjerni podaci koji su van područja ostalih podataka

Odlučiti mogu li se *outlieri* ignorirati ili oni indiciraju fenomen koji u modelu moramo uzeti u obzir.

U našem primjeru, mogu se zanemariti (radi se o nenamjernim kretnjama tijela tijekom snimanja podataka).

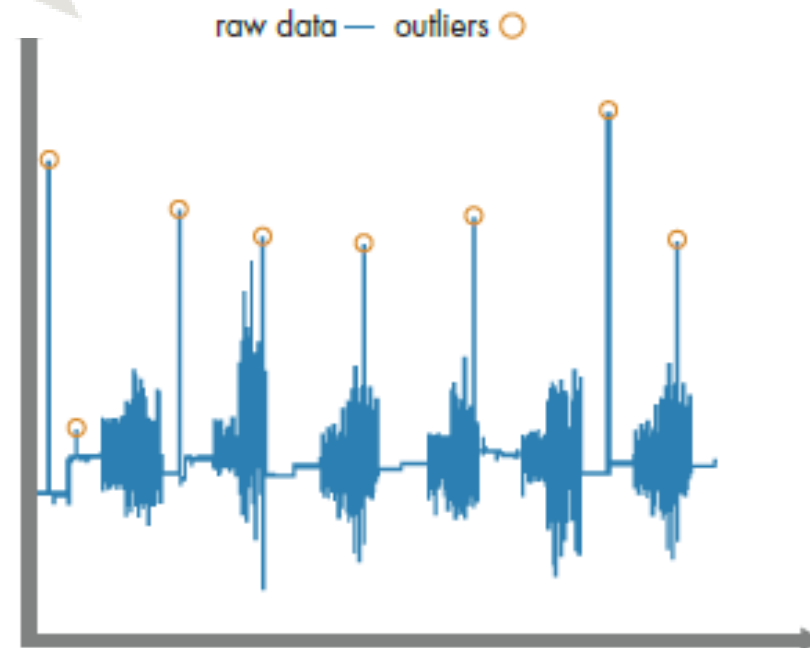
2. **Provjeriti ima li podataka koji nedostaju** (možda smo izgubili dio podataka jer je veza pala tijekom snimanja)

Nedostajuće vrijednost možemo jednostavno zanemariti, ali to će smanjiti veličinu skupa podataka.

Također možemo nedostajuće vrijednosti nadomjestiti interpolacijom ili sličnim podacima iz drugog skupa.

KORAK 2: ...Predobrada podataka...

In many applications, outliers provide crucial information. For example, in a credit card fraud detection app, they indicate purchases that fall outside a customer's usual buying patterns.



Outliers in the activity-tracking data.

KORAK DVA: ... Predobrada podataka

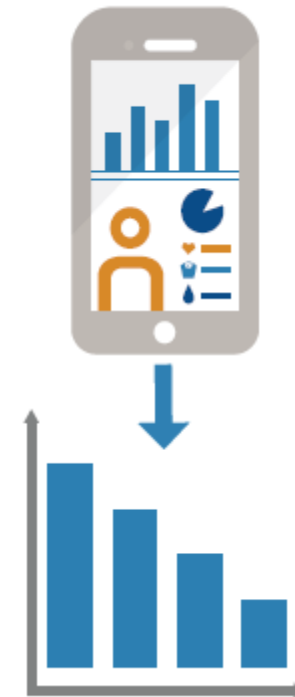
3. **Ukloniti gravitacijski efekt** iz podataka akcelerometra.

Naš algoritam treba se fokusirati na gibanje subjekta, ne gibanje telefona. Za to se primjenjuje visokopropusni filter.

4. **Podijeliti podatke** u dva skupa.

Dio podataka spremićemo za testiranje (skup za testiranje), a ostatak ćemo upotrijebiti za izgradnju modela (skup za uvježbavanje).

By testing your model against data that wasn't used in the modeling process, you see how it will perform with unknown data.



KORAK 3: Izvlačenje obilježja

DERIVE FEATURES

- Izvlačenje obilježja (**feature engineering** ili **feature extraction**) jedno je od najvažnijih koraka kod strojnog učenja;
- Njime se sirovi podaci pretvaraju u **informaciju** koju algoritam strojnog učenja može koristiti;
- Za praćenje aktivnosti želimo izvući obilježja koja “hvataju” **frekvencijski sadržaj** iz podataka akcelerometra;
- Te značajke pomažu algoritmu razlikovati **hodanje** (niske frekvencije) od **trčanja** (visoke frekvencije);
- Na temelju toga kreiramo tablicu koja uključuje odabrane značajke.



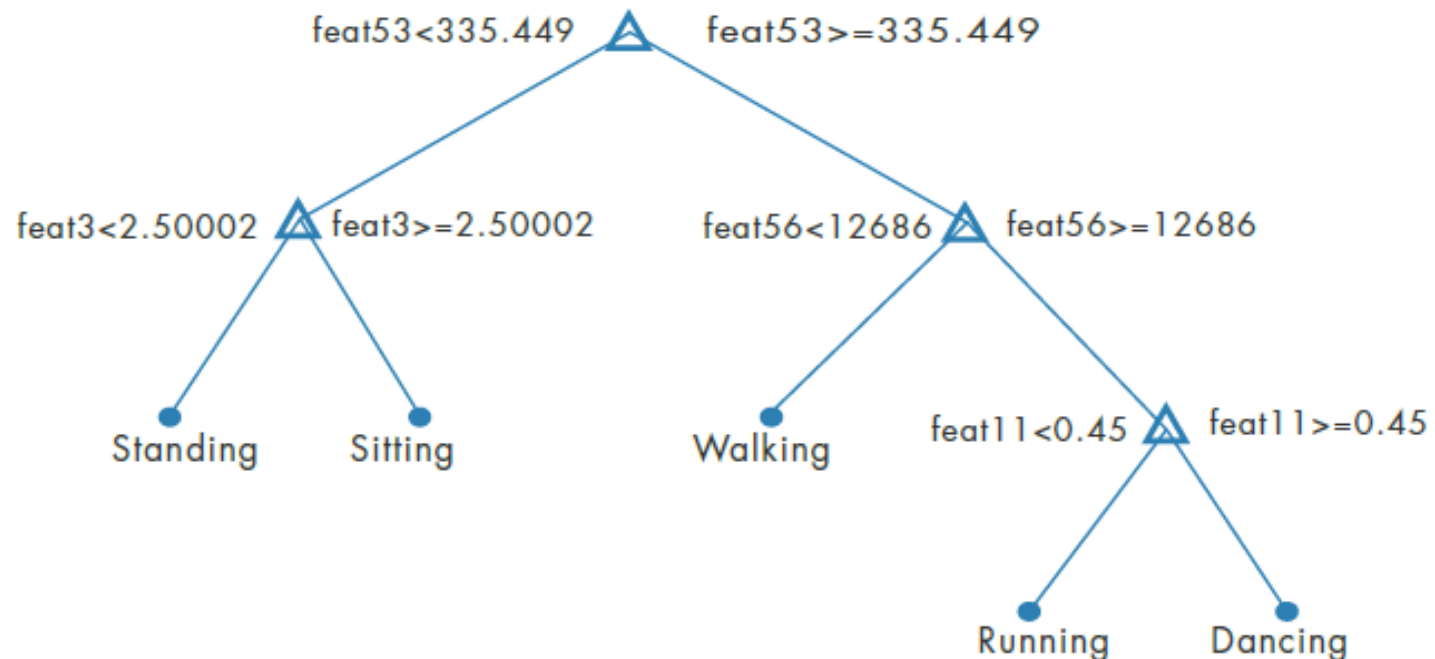
KORAK 3: ... Izvlačenje obilježlja

- Broj obilježja koje možemo izvući nije ograničen;
- Postoje standardne tehnike za različite tipove podataka.

| Data Type | Feature Selection Task | Techniques |
|-----------------------------|---|--|
| Sensor data | Extract signal properties from raw sensor data to create higher-level information | <p>Peak analysis – perform an fft and identify dominant frequencies</p> <p>Pulse and transition metrics – derive signal characteristics such as rise time, fall time, and settling time</p> <p>Spectral measurements – plot signal power, bandwidth, mean frequency, and median frequency</p> |
| Image and video data | Extract features such as edge locations, resolution, and color | <p>Bag of visual words – create a histogram of local image features, such as edges, corners, and blobs</p> <p>Histogram of oriented gradients (HOG) – create a histogram of local gradient directions</p> <p>Minimum eigenvalue algorithm – detect corner locations in images</p> <p>Edge detection – identify points where the degree of brightness changes sharply</p> |
| Transactional data | Calculate derived values that enhance the information in the data | <p>Timestamp decomposition – break timestamps down into components such as day and month</p> <p>Aggregate value calculation – create higher-level features such as the total number of times a particular event occurred</p> |

KORAK 4: Izgradnja i treniranje modela...

- Pri razvoju modela dobro je **početi** s nečim **jednostavnim** – bit će brže i lakše za interpretaciju;
- Počnimo s osnovnim stablom odlučivanja (**decision tree**):



KORAK 4: ...Izgradnja i treniranje modela...

- Za ocjenu modela nacrtajmo matricu konfuzije (**confusion matrix**) – tablicu koja uspoređuje klasifikacije učinjene modelom sa stvarnim koje smo koje smo kreirali u koraku 1:

| | | | | | |
|-------------------|------------------------|-----------------|----------------|----------------|----------------|
| | Sitting | Standing | Walking | Running | Dancing |
| TRUE CLASS | | | | | |
| Sitting | >99% | <1% | | | |
| Standing | <1% | 99% | <1% | | |
| Walking | | <1% | >99% | <1% | |
| Running | | | 1% | 93% | 5% |
| Dancing | | <1% | <1% | 40% | 59% |
| | <i>Sitting</i> | <i>Standing</i> | <i>Walking</i> | <i>Running</i> | <i>Dancing</i> |
| | PREDICTED CLASS | | | | |

- Matrica konfuzije pokazuje da naš model ima problema s razlikovanjem između trčanja i plesanja;
- Pretpostavljamo da možda stablo odlučivanja ne radi dobro za ovaj tip podataka. Probajmo nekoliko različitih algoritama.

KORAK 4: ...Izgradnja i treniranje modela...

- Pokušajmo s **K-nearest neighbors (KNN)**;
- To je jednostavni algoritam koji pohranjuje sve podatke za uvježbavanje, uspoređuje nove točke s podacima za uvježbavanje i daje najčešću klasu od “K” najbližih točaka;
- Ovom metodom dobili smo točnost od **98 %** u usporedbi s **94,1%** kod stabla odlučivanja;
- Matrica konfuzije, također, izgleda bolje.

| | | | | | |
|-------------------|------------------------|----------|---------|---------|---------|
| | Sitting | Standing | Walking | Running | Dancing |
| TRUE CLASS | Sitting | Standing | Walking | Running | Dancing |
| Sitting | >99% | <1% | | | |
| Standing | 1% | 99% | 1% | | |
| Walking | | 2% | 98% | | |
| Running | | <1% | 1% | 97% | 1% |
| Dancing | | 1% | 1% | 6% | 92% |
| | PREDICTED CLASS | | | | |

KORAK 4: ... Izgradnja i treniranje modela

- KNN treba dosta memorije za izvođenje budući da za predikciju uzima sve podatke za uvježbavanje;
- Na kraju smo pokušali s metodom **support vector machine (SVM)** i dobili vrlo dobre rezultate – **99 %** točnosti.

| | | | | | | |
|------------|----------|-----------------|----------|---------|---------|---------|
| | | Sitting | Standing | Walking | Running | Dancing |
| TRUE CLASS | Sitting | >99% | <1% | | | |
| | Standing | <1% | >99% | <1% | | |
| | Walking | | <1% | >99% | | |
| | Running | | | <1% | 98% | 2% |
| | Dancing | | <1% | <1% | 3% | 96% |
| | | Sitting | Standing | Walking | Running | Dancing |
| | | PREDICTED CLASS | | | | |

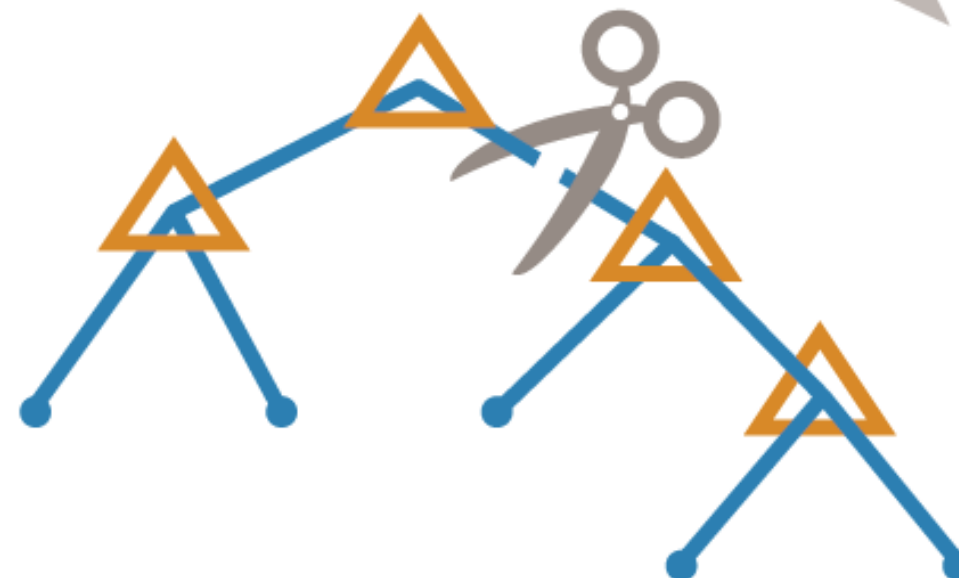
- Naš cilj ostvarili smo iteracijom modela i primjenom različitih algoritama;
- Ako još uvijek nismo zadovoljni razmotrit ćemo načine za **poboljšanje modela**.

KORAK 5: Poboljšanje modela

- Poboljšanje modela provodi se na jedan od dva načina:
 - pojednostaviti model
 - dodati kompleksnost modelu

A good model includes only the features with the most predictive power. A simple model that generalizes well is better than a complex model that may not generalize or train well to new data.

In machine learning, as in many other computational processes, simplifying the model makes it easier to understand, more robust, and more computationally efficient.

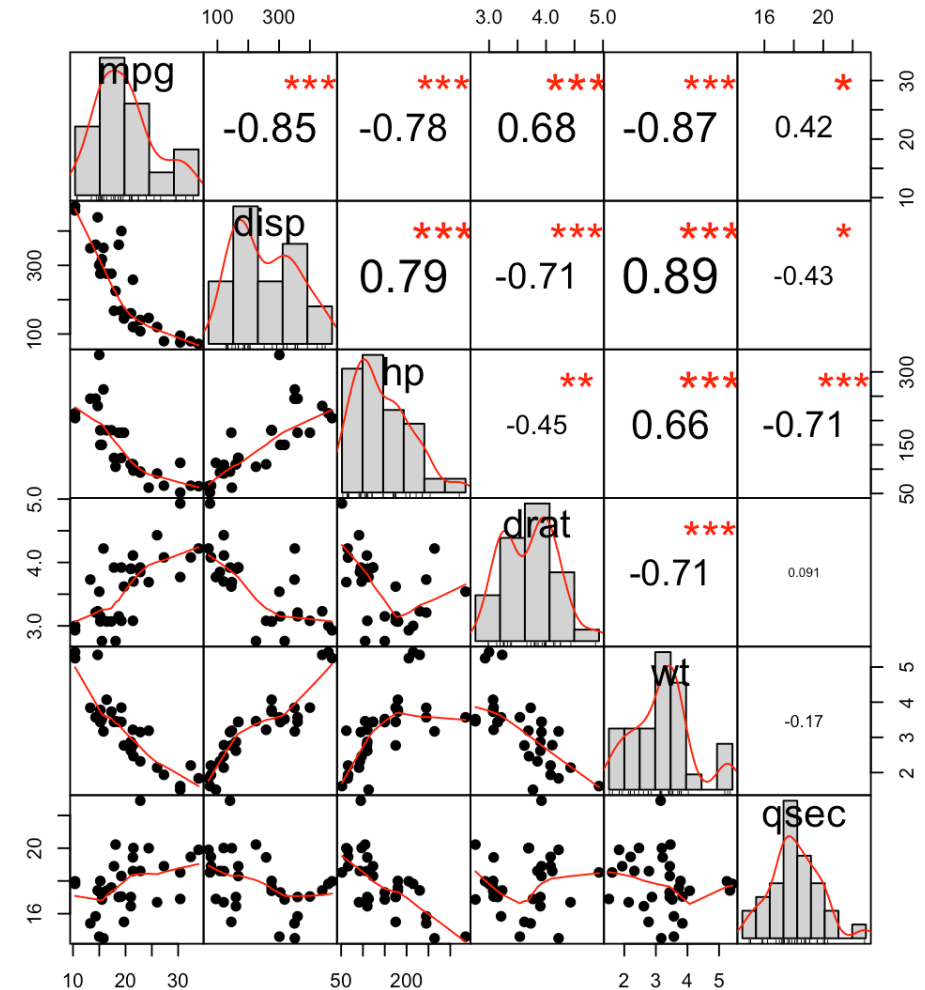


KORAK 5: Poboljšanje modela - Pojednostavljenje

- Za smanjenje broja obilježja popularne su sljedeće tehnike:

Matrica korelacije (Correlation matrix)

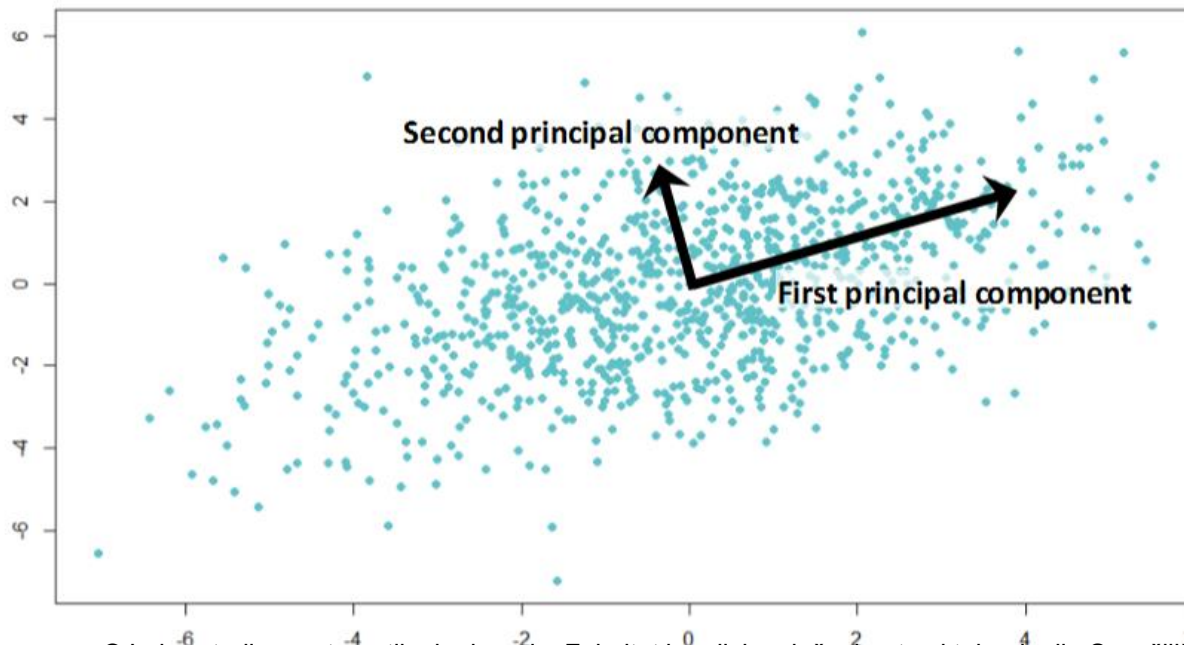
Prikazuje relaciju između varijabli pa se varijable (ili obilježja) koja nisu jako korelirana mogu ukloniti;



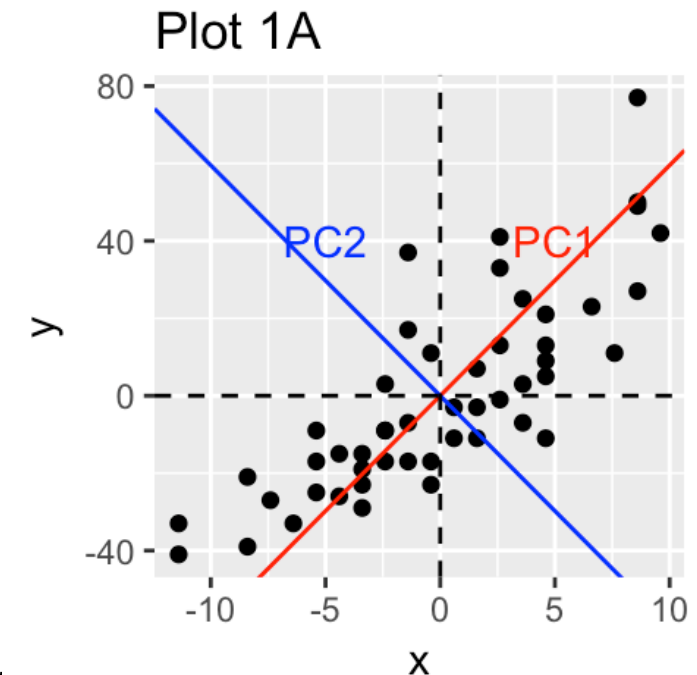
KORAK 5: Poboljšanje modela - Pojednostavljenje

Analiza glavnih komponentata (Principal component analysis – PCA)

Linearna transformacija podataka koja većinu varijance visokodimenzijskih skupova opisuje s prvih nekoliko glavnih komponentata. Prva opisuje većinu varijance, zatim druga, itd.



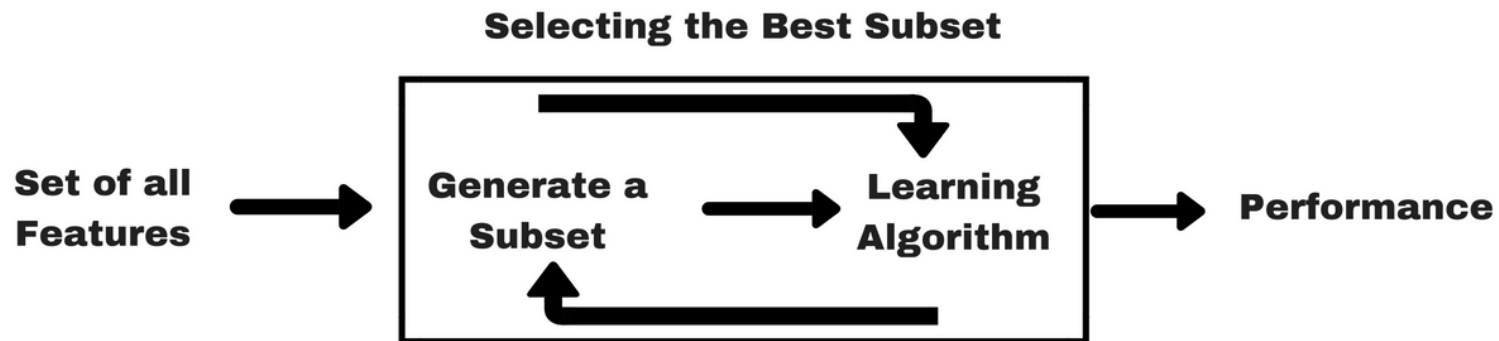
© Laboratorij za automatiku i mjerenja, Fakultet kemijskog inženjerstva i tehnologije Sveučilišta u Zag



KORAK 5: Poboljšanje modela - Pojednostavljenje

Sekvencijalna redukcija obilježja (**Sequential feature reduction**)

Reducira obilježja modela iteracijski sve dok više nema poboljšanja u radu.



KORAK 5: Poboljšanje modela – Povećanje kompleksnosti

- Ako naš model ne može razlikovati npr. šetnju od trčanja jer je pretjerano poopćen (**overgeneralised**), treba naći način za fino ugađanje:

Kombinirati modele

Spajanje **više jednostavnijih modela u veći model** koji može bolje modelirati trendove u podacima nego što bi to radili pojedini jednostavniji modeli.

Uzeti više izvora podataka

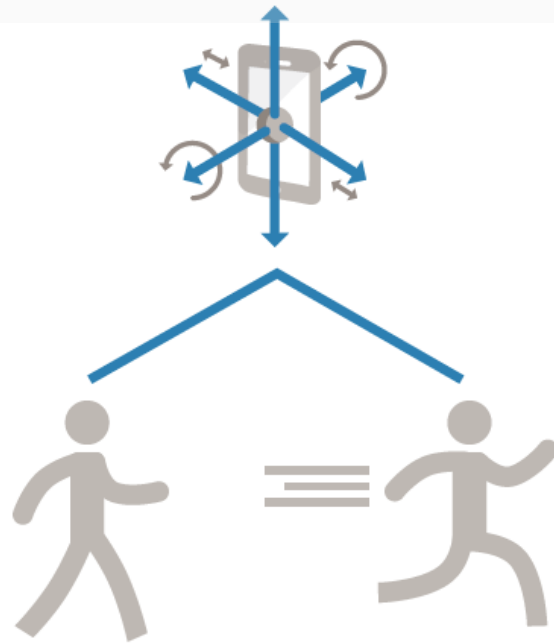
Pogledajte podatke sa žiroskopa i akcelerometra.

Žiroskop bilježi **orijentaciju mobitela** tijekom aktivnosti.

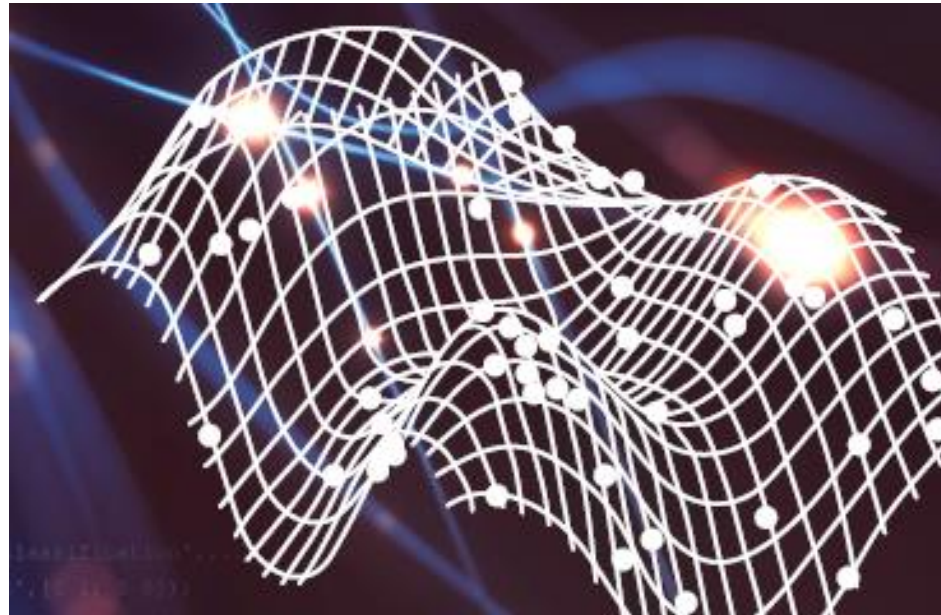
Ovaj podatak može pružiti jedinstvenu signaturu za različite aktivnosti; npr. postoji kombinacija ubrzanja i rotacije koja je karakteristična za trčanje.

KORAK 5: Poboljšanje modela – Povećanje kompleksnosti

- Kad podesimo model, ocjenjujemo njegov rad na **skupu za testiranje** kojeg smo pripremili tijekom predobrade;
- Ako model može pouzdano klasificirati aktivnosti na testnim podacima, **spreman je za primjenu** u mobilnom telefonu pa možemo početi pratiti naše aktivnosti.



PRIMJENA NENADZIRANOG UČENJA



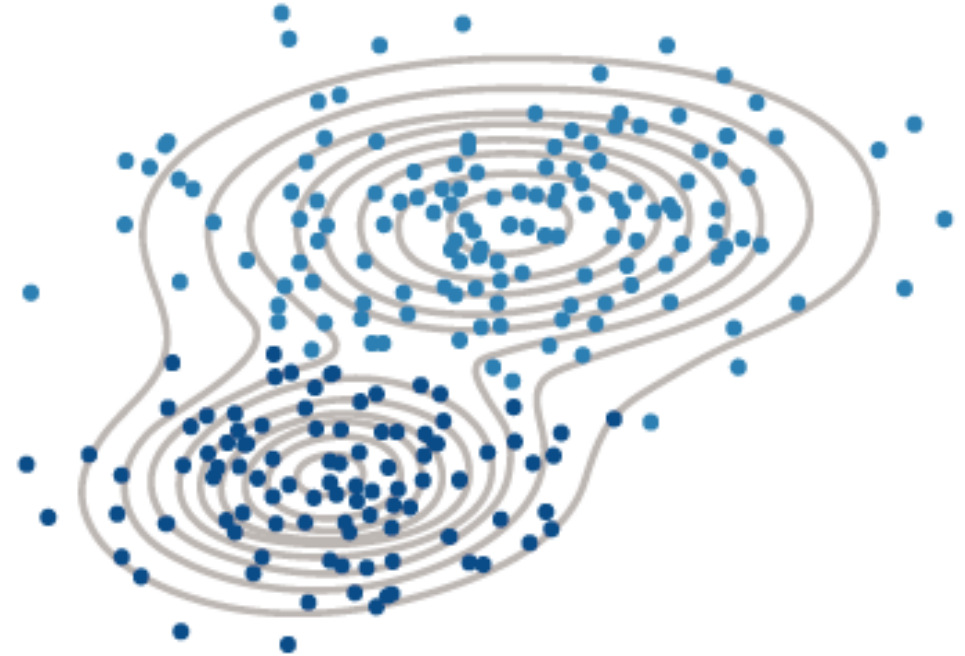
- Nenadzirano učenje je korisno kad se želi istraživati podatke dok još ne postoji specifični cilj ili nismo sigurni koje informacije sadrže podaci;
- Također je to dobar način reduciranja dimenzije podataka.

TEHNIKE NENADZIRANOG UČENJA

- Većina ovih tehnika spada u jedan od oblika analize klastera (**cluster analysis**);
- Kod analize klastera podaci se dijele u grupe na temelju neke **mjere sličnosti** ili **zajedničkih karakteristika**;
- Objekti u istom klasteru vrlo su slični, a objekti u različitim se znatno razlikuju;
- Dvije velike grupe ovih algoritama su:

Hard clustering – svaki podatak pripada samo jednom klasteru,

Soft clustering – svaki podatak može pripadati u dva ili više klastera.



Gaussian mixture model used to separate data into two clusters.

If you don't yet know how the data might be grouped:

- Use self-organizing feature maps or hierarchical clustering to look for possible structures in the data.
- Use cluster evaluation to look for the "best" number of groups for a given clustering algorithm.

„HARD CLUSTERING” ALGORITMI

k-Means

- Dijeli podatke u k međusobno isključivih klastera;
- Koliko se dobro točka uklapa u klaster određuje se shodno udaljenosti te točke od središta klastera;
- **Primjena:** kad je poznat broj klastera;
- Za brzo grupiranje velikih skupova podataka.



k-Medoids

- Slično *k-means*, ali s uvjetom da se centri klastera podudaraju s točkama podataka;
- **Primjena:** kad je poznat broj klastera
- Za brzo grupiranje kategorijskih podataka;
- Za velike skupove podataka.

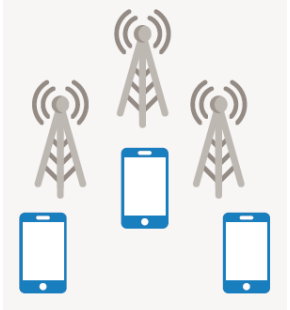


PRIMJER „HARD CLUSTERING” ALGORITMA...

Primjer...

Primjena *k-Means Clustering*-a za određivanje položaja stupova mobilne telefonije

- Tvrtnica želi znati **broj** i **položaj** stupova za mobilnu telefoniju koji će pružiti najpouzdaniju uslugu;
- Za optimalan prijam signala tornjevi moraju biti smješteni unutar klastera ljudi;
- Projekt počinje s početnom pretpostavkom o potrebnom broju klastera;
- Za procjenu inženjeri uspoređuju rad s tri i četiri tornja da ustanove kako dobro mogu klasterirati (drugim riječima, koliko dobar signal pružaju tornjevi).



... PRIMJER „HARD CLUSTERING” ALGORITMA

...Primjer

- Telefon može komunicirati samo s jednim tornjem – radi se o **hard clustering**-u;
- **k-means** tretira svaku opservaciju u podacima kao objekt koji ima svoju lokaciju u prostoru;
- Zapravo pronalazi podjelu u kojoj su objekti unutar svakog klastera što je moguće bliže jedni drugima i što je moguće dalje od objekata u drugim klasterima;
- Nakon pokretanja algoritma, moguće je odrediti rezultate podjele podataka u tri i četiri klastera.

