#### Strojno učenje 1

## 1. Uvod u strojno učenje

prof. dr. sc. Jan Šnajder prof. dr. sc. Bojana Dalbelo Bašić

Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva

Ak. god. 2022./2023.

# Sadržaj

1 Što i zašto

2 Tri pristupa i pet škola

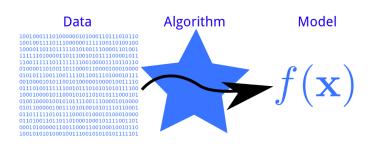
Organizacija predmeta

# Motivacija

- "We are drowning in data, but we are starved for knowledge"
   John Naisbitt, Megatrends
- Podatci = sirove informacije
- Znanje = uzorci u podatcima ili modeli koji ih objašnjavaju



## Strojno učenje



 Algoritmi za (polu)automatsku ekstrakciju novog i korisnog znanja – u obliku pravila, uzoraka ili modela – iz proizvoljnih skupova podataka

## Strojno učenje

### Strojno učenje (Alpaydin 2009)

Strojno učenje jest programiranje računala na način da optimiziraju neki kriterij uspješnosti temeljem podatkovnih primjera ili prethodnog iskustva.

- Raspolažemo modelom koji je definiran do na neke parametre
- Učenje: optimizacija parametara modela temeljem podataka
- Na temelju viđenih podataka, model mora moći predvidjeti svojstva novih, još neviđenih podataka
- Cilj strojnog učenja: izgraditi modele koji dobro generaliziraju

## Interdisciplinarnost strojnog učenja

- Umjetna inteligencija
- Teorija vjerojatnosti
- Statistika
- Optimizacija
- Teorija informacije
- Teorijska računarska znanost
- Baze podataka
- Pretraživanje informacija (engl. information retrieval)
- Psihologija, neurobiologija, kognitivna znanost

## Svi putevi vode k...strojnom učenju

#### Matematika: statistika

17. stoljeće: Vjerojatnost

18. stoljeće: Prvi Bayesovski pristupi

20. stoljeće: Frekventisti

▶ 1980.: Značajan razvoj Bayesovskih pristupa (MCMC)

#### • Elektrotehnika: Obrada signala

Početci u drugome svjetskom ratu

▶ 1960.–1970.: FFT, projektiranje filtara

▶ 1980.–1990.: Očitavanje signala, obrada slike

## Svi putevi vode k...strojnom učenju

#### Računarstvo: UI

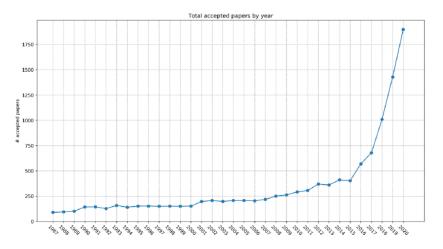
- ▶ 1940.: Perceptron
- ▶ 1960.–1970.: Logičko programiranje
- ▶ 1980.: Backpropagation, PAC, Bayesove mreže
- 2000.+: SVM, PGM, Transfer learning, Structured learning, Online learning, Deep learning, . . .

#### • Računarstvo: Razne komplementarne ideje

- Baze podataka: Dubinska analiza podataka, statističke baze podataka (OLAP), probabilističke baze podataka
- Pretraživanje informacija: PageRank
- Računalni sustavi: MapReduce, Hadoop

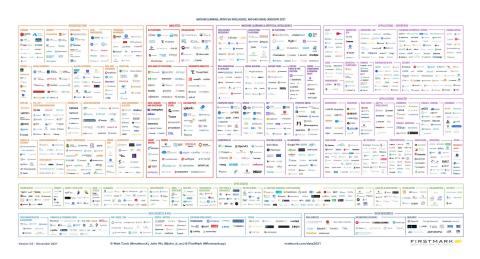
# Strojno učenje danas

#### Broj radova na konferenciji NeurIPS



https://towardsdatascience.com/neurips-conference-historical-data-analysis-e45f7641d232

## Krajolik alata strojnog učenja



https://mattturck.com/data2021/

## Primjene strojnog učenja

- Složeni problemi ne postoji ljudsko znanje o procesu ili ljudi ne mogu dati objašnjenje procesa (npr. raspoznavanje govora)
  - problemi koje nije moguće riješiti na klasičan algoritamski način (UI-potpuni problemi)
- Sustavi koji se dinamički mijenjaju
- Ogromne količine podataka ima li znanja u njima?
  - otkrivanje znanja u skupovima podataka (engl. data mining)

# Slučaj 1: Teški problemi

- Raspoznavanje uzoraka
- Računalni vid
- Obrada prirodnog jezika
- Raspoznavanje govora
- Robotika
- . . .



# Slučaj 2: Dinamički prilagodivi sustavi

- Inteligentni računalni sustavi moraju se moći prilagoditi svojoj okolini
- Sustav koji je sposoban učiti može se bolje prilagoditi novim situacijama



- Robotika
- Višeagentni sustavi
- Inteligentna korisnička sučelja
- ...ali zapravo bilo kakav inteligentni agent (računalni program koji djeluje u pravoj ili stvarnoj okolini, npr., virtualni asistent, algoritam za igru Go, dron za spašavanje...)

## Slučaj 3: Od podataka do znanja

- Podatci  $\rightarrow$  informacije  $\rightarrow$  znanje  $\rightarrow$  odluka
- Podataka ima u izobilju (web, tekst, eksperimentalni podatci, skladišta podataka deep web, logovi)
- S druge strane, znanja nema puno i ono je skupo
- Cilj: izgradnja modela koji objašnjavaju podatke i omogućavaju zaključivanje/predviđanje

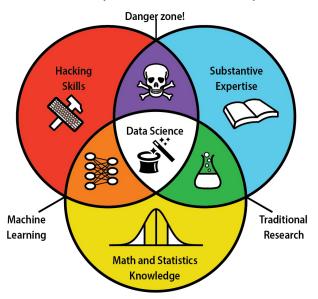


## Dubinska analiza podataka

**Dubinska analiza podataka** (engl. *data mining*) ili **otkrivanje znanja u skupovima podataka** (engl. *knowledge discovery in datasets*) – primjena strojnog učenja na (polu)strukturirane baze podataka

- Prodaja: analiza potrošačke košarice, CRM
- Financije: određivanje kreditne sposobnosti, detekcija zlouporaba kartica
- Proizvodnja: optimizacija, troubleshooting
- Medicina: postavljanje dijagnoza, prevencija
- Telekomunikacije: optimizacija usluga
- Bioinformatika: analiza izražajnosti gena, poravnavanje gena
- Text mining: klasifikacija teksta, ekstrakcija informacija
- ...

## Znanost o podatcima (engl. Data Science)



# Sadržaj

1 Što i zašto

2 Tri pristupa i pet škola

Organizacija predmeta

## Tri pristupa strojnom učenju

## (1) Nadzirano učenje (engl. supervised learning)

Podatci su parovi (ulaz, izlaz)=(x,y), treba pronaći  $\hat{y}=f(x)$ 

- ullet Ako je y je diskretna/nebrojčana vrijednost: klasifikacija
- ullet Ako je y kontinuirana/brojčana vrijednost: regresija

## (2) Nenadzirano učenje (engl. unsupervised learning)

Dani su podatci bez ciljne vrijednosti, treba naći pravilnost u podatcima

- grupiranje (engl. clustering)
- procjena gustoće (engl. density estimation)
- smanjenje dimenzionalnosti (engl. dimensionality reduction)

## (3) Podržano/ojačano učenje (engl. reinforcement learning)

Učenje optimalne strategije na temelju pokušaja s odgođenom nagradom

## Primjena klasifikacije: predviđanje ishoda izbora

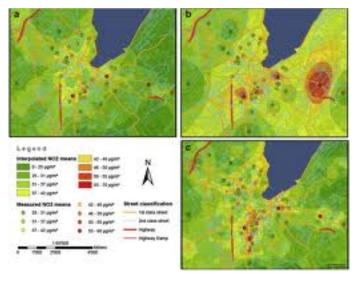
#### 2016 Presidential Election Prediction

A Predictive-Descriptive Artificial Intelligence-Based Expert Computer System



Election Analytics (https://www.experfv.com/uploads/ckeditor/pictures/132/content NikrayeshSlide3.JPG)

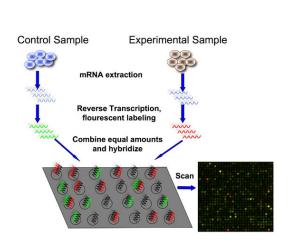
# Primjena regresije: predviđanje razine emisije NO<sub>2</sub>

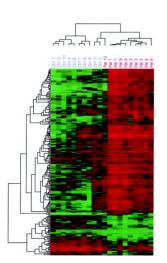


Exposure models for traffic related NO2 (http://www.sciencedirect.com/science/article/pii/S1352231011009629)

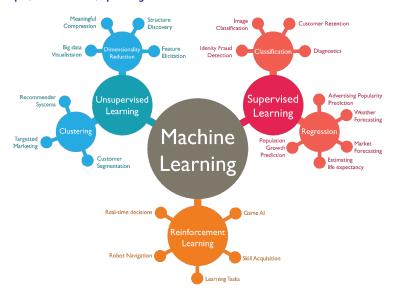
# Primjena grupiranja: analiza DNK-mikropolja

Cilj: grupiranje gena sa sličnom izražajnošću (slična izražajnost – slična funkcionalnost)

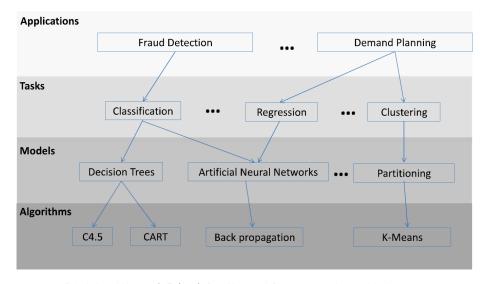




### Pristupi, zadatci, primjene



## Primjene, zadatci, modeli i algoritmi



Rokach, L. and Maimon, O. Z. (2008). Data Mining with Decision Trees: Theory and Applications.

## Pet škola strojnog učenja

Pedro Domingos: The Master Algorithm



Škola	Temelji	Glavni algoritam
Symbolists	Logika, filozofija	Indukcija
Connectionists	Neuroznanost	Backpropagation
<b>Evolutionaries</b>	Evolucijska biologija	Genetičko programiranje
Bayesians	Statistika	Probabilističko zaključivanje
Analogizers	Psihologija	Jezgreni strojevi

# Pet škola strojnog učenja

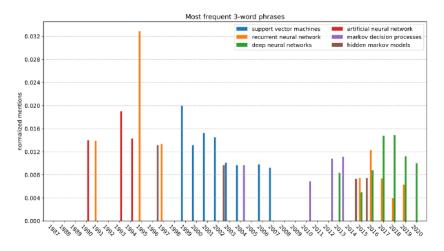
Pedro Domingos: The Master Algorithm



Škola	Cilj	
Symbolists	Popunjavanje rupa u znanju	
Connectionists	Emulacija mozga	
<b>Evolutionaries</b>	Simulacija evolucije	
Bayesians	Sustavno smanjivanje neizvjesnosti	
Analogizers	Uočavanje sličnosti između staroga i novoga	

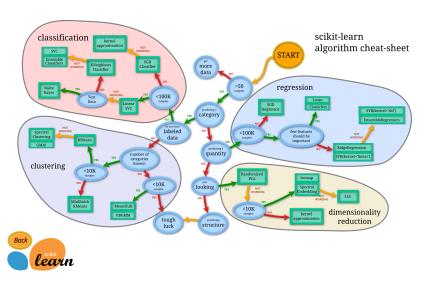
#### Trendovi

#### Najčešće fraze u radovima na konferenciji NeurIPS



https://towardsdatascience.com/neurips-conference-historical-data-analysis-e45f7641d232

## Karta strojnog učenja



https://scikit-learn.org/stable/tutorial/machine\_learning\_map/

### Još primjera...

Jupyter-bilježnica 1: Uvodni primjer https://goo.gl/0Vd26Z

Svi primjeri dostupni su na:

https://github.com/jsnajder/StrojnoUcenje

# Sadržaj

1) Što i zašto

2 Tri pristupa i pet škola

Organizacija predmeta

#### Izvođači

#### Predavač:

• Prof. dr. sc. Jan Šnajder

#### Asistenti:

- Zoran Medić, mag. ing. (glavni asistent)
- Ana Barić, Ivana Bašljan Puceković, Ivan Bilić, Filip Karlo Došilović, David Dukić, Antea Hadviger, Fran Jelenić, Josip Jukić, Miha Keber, Josipa Lipovac, Domagoj Pluščec

#### Demosi:

 Dario Deković, Mladen Džida, Ivan Furač, Josip Hrvatić, Helena Ladić, Ivan Martinović, Mirta Moslavac, Magdalena Mucić, Leon Novački, Luka Pavlović, Lovre Petešić, Ivan Rep, Matej Škrabić, Matea Vasilj, Janko Vidaković

## O predavaču

- Jan Šnajder, PhD
- http://www.zemris.fer.hr/~jan/
- Diplomirao (2002) i doktorirao (2010) računarsku znanost na FER-u
- Poslijedoktorand na Sveučilištu u Heidelbergu, znanstveno usavršavanje na Sveučilištu u Stuttgartu, NICT Kyoto, Sveučilištu u Melbourneu
- Član Laboratorija za analizu teksta i inženjerstvo znanja (TakeLab)
- Istraživački interesi: obrada prirodnog jezika (NLP) i strojno učenje (ekstrakcija informacija, tekstna analitika, leksička seminantika, analiza subjektivnog teksta i primjene u društvenim znanostima)

### O nama



Text Analysis and Knowledge Engineering Lab <a href="http://takelab.fer.hr">http://takelab.fer.hr</a>

#### Plan nastave

### 23 teme organizirane u 6 cjelina:

- (1) Uvod u strojno učenje, Osnovni koncepti, Linearna regresija
- (2) Linearni diskriminativni modeli, Logistička regresija
- (3) Stroj potpornih vektora, Jezgrene metode, Neparametarske metode
- (\*) Opcionalno: Ansambli Međuispit
- (4) Procjena parametara, Bayesov klasifikator
- (5) Probabilistički grafički modeli
- (6) Grupiranje, Vrednovanje modela
- (\*) Opcionalno: Odabir značajki
  - Završni ispit

### Ishodi učenja

- Definirati osnovne pojmove strojnog učenja
- Objasniti teorijske pretpostavke, matematičke osnove te prednosti i nedostatke temeljnih algoritama (ne)nadziranog strojnog učenja
- Primijeniti temeljne algoritme klasifikacije, regresije i grupiranja na jednostavnije probleme
- Primijeniti postupak odabira i evaluacije točnosti modela
- Analizirati i usporediti temeljne algoritme strojnog učenja s obzirom na njihove komponente i računalnu složenost
- Razlikovati temeljne pristupe strojnom učenju (generativni/diskriminativni, parametarski/neparametarski, bayesovski/frekventistički) te objasniti poveznice
- Procijeniti prikladnost temeljnih algoritama strojnog učenja za zadani zadatak
- Oblikovati i implementirati jednostavnije algoritme za klasifikaciju, regresiju i grupiranje

### U konačnici, želimo da...

- Razumijete osnovne ideje i razine apstrakcije
- Razumijete tipičnu matematiku ispod toga
- Shvatite algoritme i metode tako da ih implementirate ili ispitate na jednostavnim primjerima
- 4 Izgradite svoju kartu područja i dobijete putokaze za samostalno učenje

# Dajemo strukturu, detalje popunjavate sami...



#### Nastavne aktivnosti

#### Predavanja/vježbe:

- HR: ponedjeljkom 10-12 sati i srijedom 8-10 sati u B4
- eN: utorkom 9–11 sati u A302 i četvrtkom 12–14 sati u D152

#### Domaće zadaće:

• 6 zadaća koje predajete online

#### Laboratorijske vježbe:

4 vježbe koje predajete pa demonstrirate asistentima

### Demonstracije:

dodatni termini u kojima demosi rješavaju zadatke s ispita

#### Konzultacije:

• Petkom 11–12 sati; prijava on-line najkasnije dva dana ranije

### Ispiti:

- Međuispit + završni ispit
- Ispitni rokovi: pismeni ispit + usmeni ispit

### Predavanja

- 20 predavanja i 6 rekapitulacija
- Klasična predavanja (ja objašnjavam na ploči, vi pratite, radite bilješke i pitate ako je nešto nejasno)
- Podijeljena u 6 cjelina, nakon svake cjeline radimo rekapitulaciju i rješavamo nekoliko zadatka s prošlogodišnjih ispita (aka auditorne vježbe)
- Kratka kviz-pitanja na satu (za motivaciju i formativnu povratnu informaciju), preko Moodlea, za ukupno 6 bonus bodova
  - Preporučamo da dolazite na predavanja, ali dolazak nije obavezan (svi materijali su vam dostupni za samostalno učenje)

# Materijali za učenje

- Videopredavanja
- 2 Skripte s predavanja (tekst predavanja s primjerima, dodatnim bilješkama i literaturom za one koji žele znati više (13)
- 3 Natuknice s predavanja (vrlo sažeta predavanja)
- 4 Zadatci za učenje i vježbu (uključivo zadatci s ispita)
- 5 Laboratorijske vježbe

Sve je javno dostupno na https://www.fer.unizg.hr/predmet/struce1/materijali

#### Dodatna literatura

 Ethem Alpaydin: *Introduction to Machine Learning*, 3rd edition, MIT Press, 2015.

Christopher Bishop:
 Pattern Recognition and Machine Learning,
 Springer, 2006.

Kevin P. Murphy: Probabilistic Machine Learning: An Introduction, MIT Press, 2022.







#### Dodatna dodatna literatura

#### Osnovno:

- Hastie, Tibshirani, Friedman: Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2003.
- Deisenroth, Faisal, Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.
- Marsland: Machine Learning: An Algorithmic Perspective, Chapman and Hall/CRC, 2009.
- Duda, Hart, Stork: Pattern Classification, Wiley-Interscience, 2000.

#### Napredno:

- Murphy: Probabilistic Machine Learning: Advanced Topics, MIT Press, 2022.
  - Mohri, Rostamizadeh, Talwalkar: Foundations of Machine Learning, MIT Press, 2012.
  - Shalev-Shwartz, Ben-David: Understanding machine learning: From theory to algorithms, Cambridge University Press, 2014.

#### Specifično:

- Koller, Friedman: Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
- Goodfellow, Bengio, Courville: Deep Learning, MIT Press, 2016.

#### Korisne web-stranice

- Data Science StackExchange
   Podatkovna znanost, analiza podataka, strojno učenje
   https://datascience.stackexchange.com/
- CrossValidated QA
   Statistika, dubinska analiza i vizualizacija podataka
   http://stats.stackexchange.com/

### Domaće zadaće

- 6 domaćih zadaća, radite ih nakon rekapitulacije, za bonus bodove
- Zadatci za vježbu koje rješavate na papiru i uploadate sken na Moodle kao jedan PDF file
- Bodovanje: maks. 1 bonus bod po zadaći, ukupno 6 bonus bodova
- Domaće zadaće rješavajte samostalno. Možete se, naravno, konzultirati s drugima, ali drugi ne smiju rješavati zadaće umjesto vas
- Domaće zadaće usklađene su s ishodima učenja predmeta, dok ispiti provjeravaju te iste ishode učenja. Ergo, ako rješavate domaće zadaće, izgledno je da ćete dobro riješiti ispit

# Laboratorijske vježbe

- 4 laboratorijske vježbe, radite ih u Pythonu u okruženju SciPy korištenjem biblioteke scikit-learn i Jupyter-bilježnice
- Svrha vježbi je stecanje dubljeg razumijevanje rada algoritama. Svrha vježbi nije stjecanje iskustva rada na stvarnom projektu strojnog učenja (za to postoje drugi predmeti)
- Rješenja demonstrirate asistentima u svom terminu demonstracije
- Bodovanje: maks. 7.5 bodova po vježbi, ukupno 30 bodova
- Promjena termina moguća je jedino u parovima studenata; molimo, ne dolazite u druge termine (iznimka: slučaj više sile)
- Trebate biti sposobni na zahtjev mijenjati programski kôd te razumijeti svu teoriju na kojima se temelje modeli/algoritmi
- Programske zadatke trebate raditi samostalno. Prepisivanje rješenja predstavlja kršenje kodeksa FER-a i podliježe sankcijama

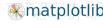
# Ekosustav SciPy

































# Laboratorijske vježbe i domaće zadaće – savjeti

- 1 Krenite raditi na vrijeme.
- Ne precjenjujte svoje sposobnosti. Nemojte krenuti raditi vježbu i zadaću tek na vikend uoči predaje.
- 3 Ako ipak niste stigli napraviti vježbu ili zadaću, nemojte je prepisati i takvu predati. Umjesto toga, naučite nešto iz toga i drugi puta krenite ranije.
- Uračunajte nepredvidive okolnosti. Možda nešto neće raditi isprve, ili će vam trebati više vremena da nešto shvatite. Ako si ostavite dovoljno vremena za takve situacije, izbjeći ćete stres.
- Ponovite teorijske osnove iza svake lab. vježbe. Svrha laboratorijskih zadataka je upravo da povežu teoriju s praksom. Asistenti će vas zato pitati i teoriju iza svakog zadatka.
- O Proučite nultu laboratorijsku vježbu, premda se ne predaje, pa čak i ako (mislite da) ste jako dobri u Pythonu.

### Ispiti

- Ispiti provjeravaju ishode učenja
- Kontinuirano: međuispit (35 bodova) + završni ispit (35 bodova)
  - gradivo obuhvaćeno međuispitom ne ulazi u završni ispit
- Ispitni rokovi: pismeni dio + usmeni dio (70 bodova)
  - usmeni ispit služi provjeri pismenog ispita i eventualnoj korekciji ocjene za granične slučajeve
  - za pristupanje usmenom ispitu potrebno je ostvariti prag od 50% bodova na pismenom ispitu
- Ispiti su obliku pitanja s ponuđenim odgovorima te sadržavaju teorijska, problemska i numerička pitanja
- Ukupno 22 pitanja, za 100% bodova dovoljno je točno riješiti njih 20
- ullet Točan odgovor nosi +1 bod, a netočan -1/3 boda
- Čuvari na ispitima ne smiju odgovarati na pitanja o ispitnim pitanjima
- Primjerci prošlogodišnjih ispita dostupni su na stranici predmeta

# Ocjenjivanje

	Kontinuirano		Ispitni rok	
	Prag	Udio	Prag	Udio
Domaće zadaće		+6%		
Laboratorijske vježbe	30%	30%	30%	30%
Kvizovi		+6%		
Međuispit		35%		
Završni ispit		35%		
Pismeni ispit			50%	35%
Usmeni ispit			50%	35%

- Prag na lab. vježbama je 9 bodova (tj. 30% od 30 bodova)
- Bonus bodovi se zbrajaju i ograničeni su na ukupno maks. 10 bodova
- Pismeni ispit nosi 35 bodova. Samo studenti koji polože pismeni ispit pristupaju usmenom ispitu

## Ocjenjivanje



Izvrstan (5)	89
Vrlo dobar (4)	76
Dobar (3)	63
Dovoljan (2)	50

Za ocjenu dovoljan potrebno je ostvariti **prag** na laboratorijskim vježbama i ukupno barem **50** bodova bez bonus bodova! Bonus bodovi pribrajaju se isključivo prolaznim ocjenama.

Ostvareni bodovi se ne zaokružuju (ukupni bodovi moraju zadovoljiti prag)

# Komunikacija

- Zbog velikog broja studenata, nažalost nismo u mogućnosti odgovarati na e-mailove
- Kontaktirajte nas putem obrasca
- Odgovor ćemo javno objaviti u dokumentu s pitanjima i odgovorima, ili ćemo vam odgovoriti mejlom (ako je pitanje specifično ili osjetljivo)
- Prije nego što putem obrasca postavite pitanje, molimo provjerite da ono već nije odgovoreno

# Vezni studenti (aka "veznjaci")

(eng. liaison students)

- Svaka tri tjedna, a po potrebi češće, iz skupa studenata koji pohađaju predavanja slučajnim ćemo mehanizmom odabrati pet studenata za ulogu "studenata veznjaka" i s njima organizirati sastanak
- Veznjaci djeluju u dvije uloge:
  - ▶ Osobna uloga: daju svoje mišljenje (o svojem napretku na predmetu, nejasnoćama, poz/neg aspektima predmeta)
  - Reprezentativna uloga: izvještavaju o problemima/idejama prikupljenima od drugih studenata. Veznjaci su nužni čuvati anonimnost studenata koje predstavljaju
- Raspravit ćemo sve probleme i promisliti eventualne intervencije u izvođenje nastave kad god je to moguće
- Ako ste odabrani kao veznjak, sudjelovanje je obavezno (osim u slučaju više sile, što trebate dojaviti) i uvjet je za polaganje predmeta. Molimo pratite obavijesti!

## U slučaju poteškoća ili nedoumica

Ako imate poteškoća sa studiranjem (probleme koje ne možete riješiti sami ili za koje trebate savjet), možete se javiti **Savjetničkoj službi FER-a**: www.fer.unizg.hr/zivot\_na\_fer-u/zdravlje\_i\_dobrobit/savjetnicka\_sluzba

Ako su problemi povezani s ovim predmetom, svakako se javite na jan.snajder@fer.hr da razmotrimo kako Vam možemo pomoći. Također, slobodno se javite ako trebate savjet ili imate stručno/karijerno pitanje u vezi strojnog učenja ili studija općenito.

## Web-stranica predmeta i materijali

- https://www.fer.unizg.hr/predmet/struce1
- https://www.fer.unizg.hr/predmet/struce1/obavijesti
- https://www.fer.unizg.hr/predmet/struce1/materijali
  - Silabus (plan predavanja)
  - Ishodi učenja (detaljno za svaku od 23 teme)
  - ▶ Pitanja i odgovori (+ poveznica na kontakt-obrazac)
  - Videopredavanja (za 18 tema)
  - Predavanja (skripte, za 18 tema)
  - Natuknice za predavanja
  - Zadatci za vježbu
  - Laboratorijske vježbe
  - Prošlogodišnji ispiti

### Sažetak

- Strojno učenje bavi se rješavanjem problema koje je teško riješiti klasičnim UI pristupima
- Podataka ima u izobilju izazov je transformirati podatke u znanje tako da možemo zaključivati i predviđati
- Strojno učenje povezano je sa statistikom, računarskom znanošću, umjetnom inteligencijom, . . .
- Učenje se svodi na optimizaciju parametara modela na temelju podataka
- Dva osnovna pristupa: nadzirano učenje (zadatci klasifikacije i regresije) i nenadzirano učenje (zadatci grupiranja, procjene gustoće, smanjenje dimenzionalnosti)



Sljedeća tema: Osnovni koncepti