Master studije
GEOINFORMATIKA

# Mašinsko učenje - definicija, osnovni pojmovi i podele

Tatjana Jakšić Krüger

tatjana@turing.mi.sanu.ac.rs

2. novembar 2021.

Građevinski fakultet Univerziteta u Beogradu

# Šta je mašinsko učenje

Cilj pohađanja ovog kursa:

- pravilno utvrditi algoritam za treniranje,
- efikasan skup vrednosti hiperpamatara,
- lakše detektovanje i otklanjanje problema.

# Šta je mašinsko učenje

Mašinsko učenje je:

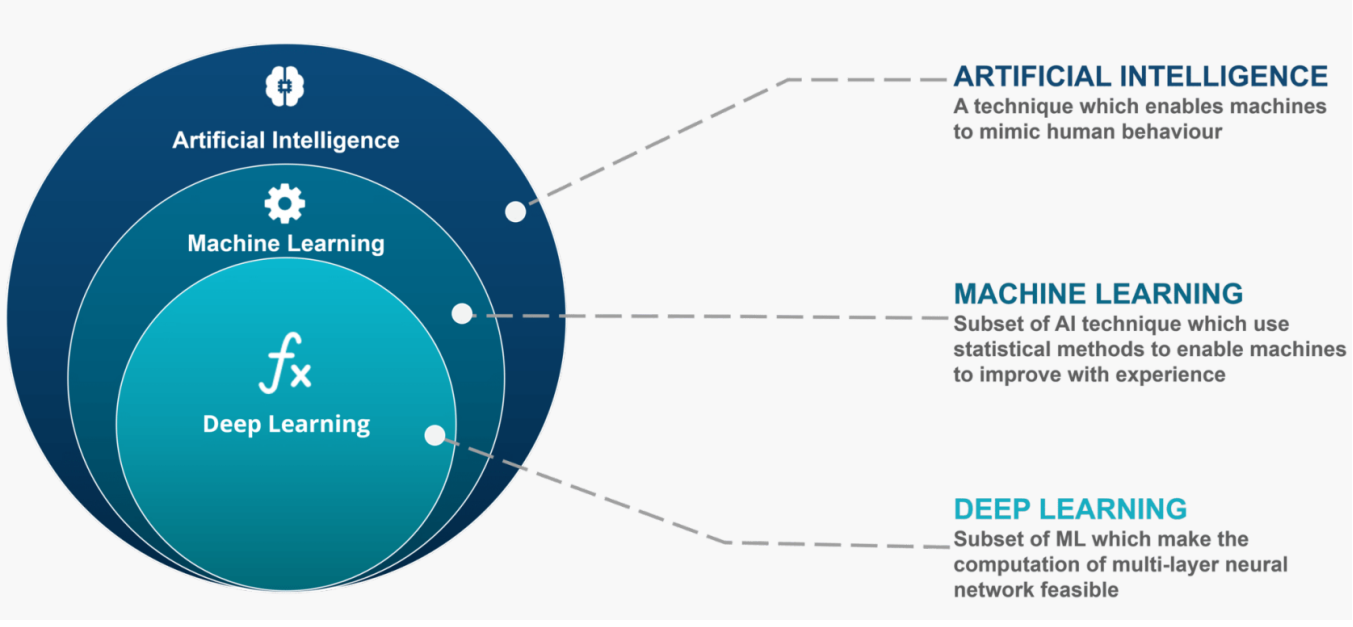
- Učenje iz podataka.
- Podgrana oblasti veštačke inteligencije (*eng.* artificial intelligence, AI). Zasnovano je na:
  - 1950-tih godina, razvoj AI.
  - 2000-tih godina, razvoj računarske moći i skladištenja podataka [1].
- Učenje i inteligencija ne opisuju iste aktivnosti.
- Učenje je sposobnost (softverskog) sistema da generalizuje na osnovu prethodnog iskustva i da odgovara na pitanja koja se tiču novih pojava/entiteta [2].
- Mašinsko učenje nije formalan metod. Ne može da dokaže tvrđenje, testira ispravnost softvera ili hardvera.

**[1]** http://ml.matf.bg.ac.rs/readings/ml.pdf

**[2]** http://ai.fon.bg.ac.rs/wp-content/uploads/2016/10/ML_intro_2016.pdf

# Šta je mašinsko učenje

Mašinsko učenje i veštačka inteligencija (Artificial Intelligence) nisu identične oblasti istraživanja.




- Ne postoji opšteprihvaćena definicija veštačke inteligencije. Može se reći da AI obuhvata svaku vrstu automatskog učenja i zaključivanja ([3] i [4]).
- Razlikujemo simbolički i statistički zasnovana veštačka inteligencija [4].

**[3]** https://www.rcc.org.rs/Vestacka_Inteligencija_prvi_deo.pdf

**[4]** http://poincare.matf.bg.ac.rs/~janicic//courses/VI_B5.pdf

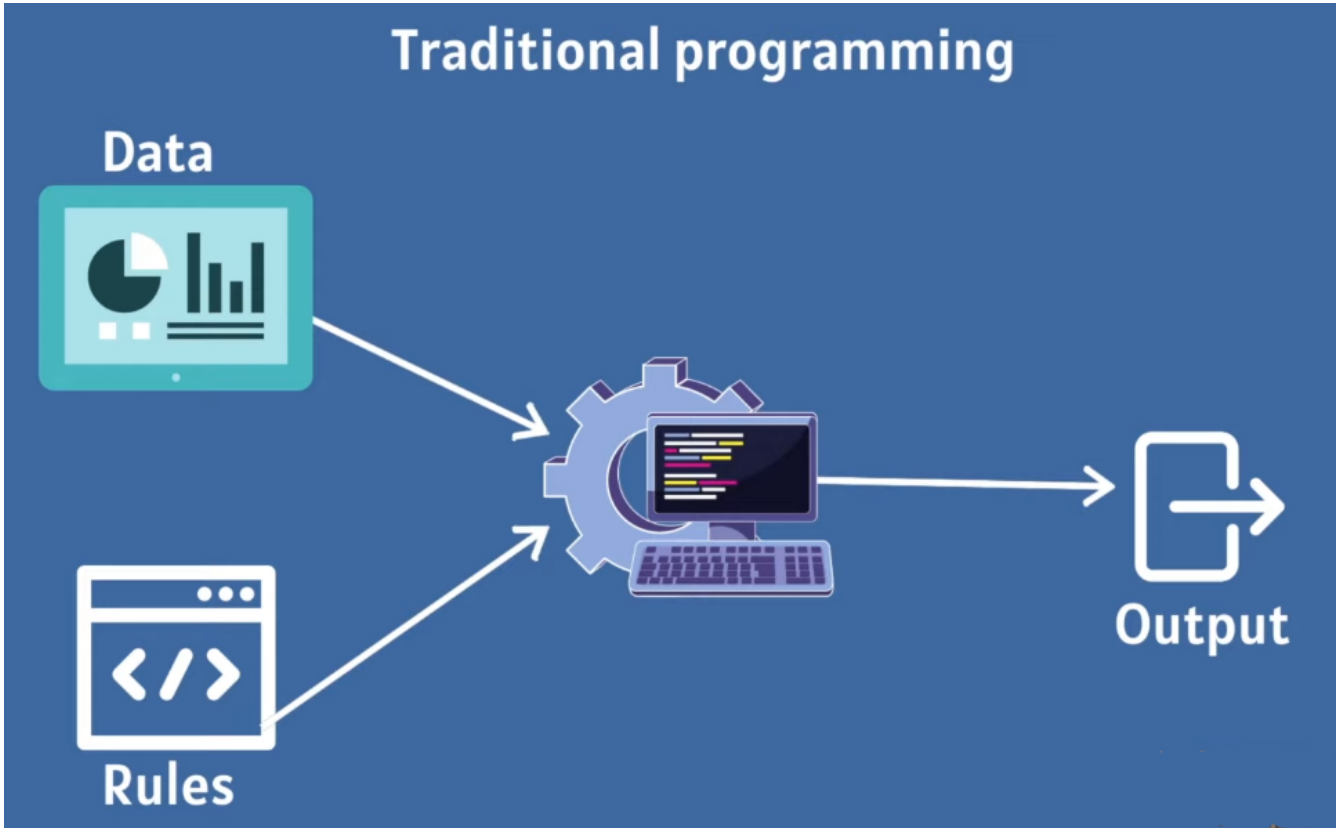
**Izvor slike**: https://ospreydata.com/ai-ml-models-101-what-is-a-model/

# Šta je mašinsko učenje

- Klasično rešavanje problema se fokusira na metodologiju.
- Računari prenose analitički račun u digitalni svet.
- Eksplicitno programiranje omogućuje tačnost i preciznost, ali na manjem broju praktičnih problema.
  - Eksplicitno programiranje podrazumeva da imamo preciznu specifikaciju problema i algoritma.
- Mašinsko učenje omogućuje rezultate za veliki broj problema iz prakse, ali sa umanjenom tačnošću i preciznošću.

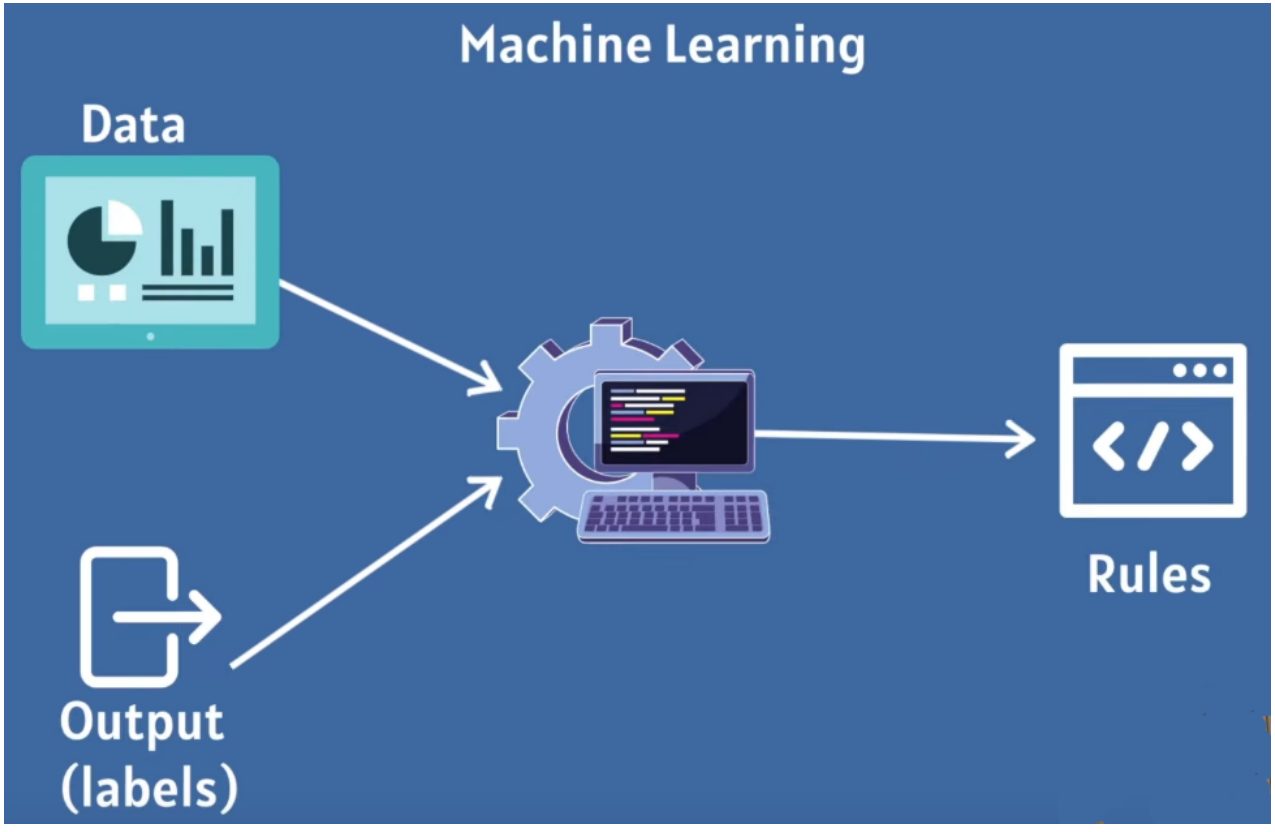
# Tradicionalno programiranje vs. mašinsko učenje

Tradicionalno (eksplicitno) programiranje vs. mašinsko učenje (implicitno programiranje).



# Tradicionalno programiranje vs. mašinsko učenje

Tradicionalno (eksplicitno) programiranje vs. mašinsko učenje (implicitno programiranje).



# Koraci mašinskog učenja

1. Prikupljanje podataka.
2. Priprema podataka.
3. Analiza rezultujućih skupova podataka.
4. Izbor jedne ili više modela učenja.
5. Obuka (training).
6. Evaluacija rafinisanih modela.
7. Konfiguracija hiperparametara (hyperparameter tuning).
8. Predviđanja.

# Motivacija

Razlog za primenu mašinskog učenja:

- rešavanje teških problema u praksi.
- Primeri:
  - prepoznavanje lica,
  - prepoznavanje razlicitih objekata i oblika na slikama i video zapisima,
  - klasifikacija teksta, mašinsko prevođenje,
  - analiza osećanja izraženih u tekstu,
  - prepoznavanje govora, itd.

# Definicija

Arthur Samuel, 1959. "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort." [5]

## Definition

Mašinsko učenje je oblast istraživanja koja omogućuje računaru da uči bez da je eksplicitno programiran.

## Definition (Tom Mitchell, 1998.)

Računarski program uči iz iskustva $I$ u zavisnosti od grupe zadataka $Z$ i mere perfomansi $P$ ukoliko njegova perfomansa nad zadatkom iz $Z$ se poboljšava zahvaljujući iskustvu $I$.

**[5]** https://ieeexplore.ieee.org/abstract/document/5389202

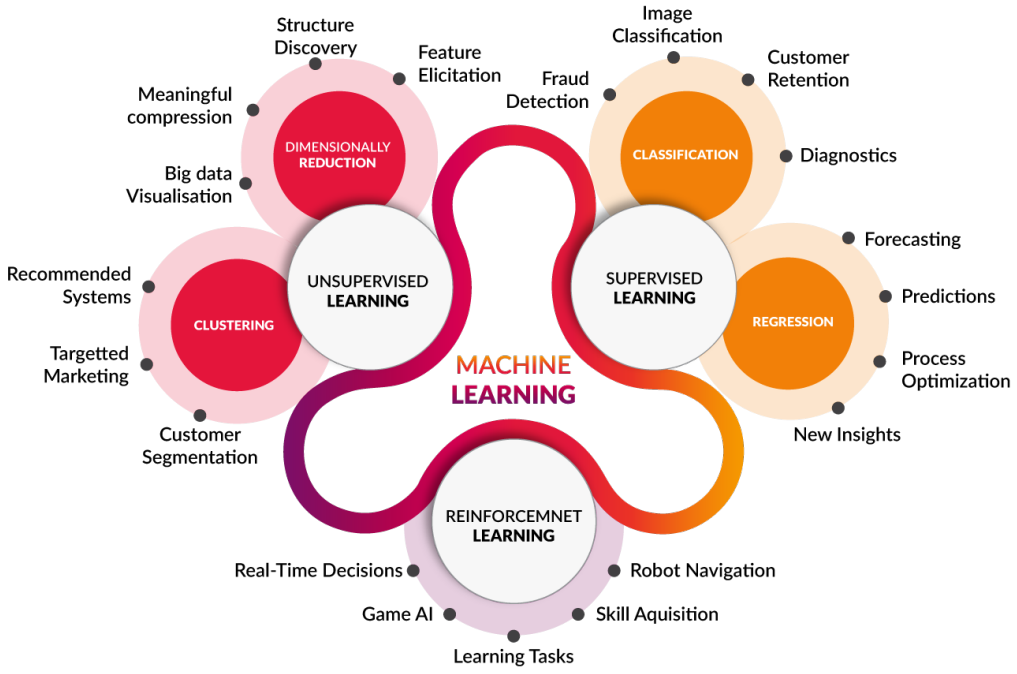
# Podela

Tipovi mašinskog učenja:

- Nadgledano učenje (eng. supervised);
- Nenadgledano učenje (eng. unsupervised);
- Učenje potkrepljivanjem ili učenje uz podsticaje (eng. reinforcement learning).

# Tipovi mašinskog učenja

- Nadgledano učenje (supervised learning):
  - Regresija (regression):
    - Predviđanja,
    - Optimizacija procesa.
  - Klasifikacija (classification)
    - Klasifikacija slika,
    - Otkrivanje prevare,
    - dijagnostikovanje, itd.

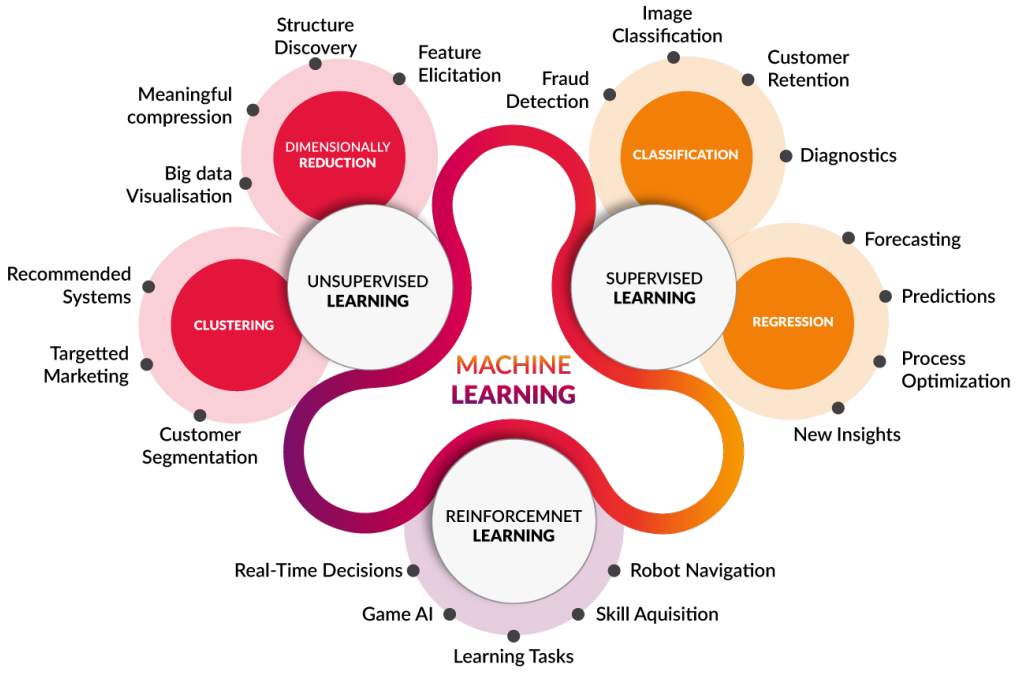



**[5] Izvor slike**:

`https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d`

# Tipovi mašinskog učenja

- Nenadgledano učenje (unsupervised learning):
  - Grupisanje (clustering):
    - Semgentacija klijenata
    - usmereno oglašavanje,
    - problem preporučivanja.
  - Smanjenje dim. problema:
    - Vizualizacija velikih podataka,
    - Otkrivanje struktura.
    - dijagnostikovanje, itd.

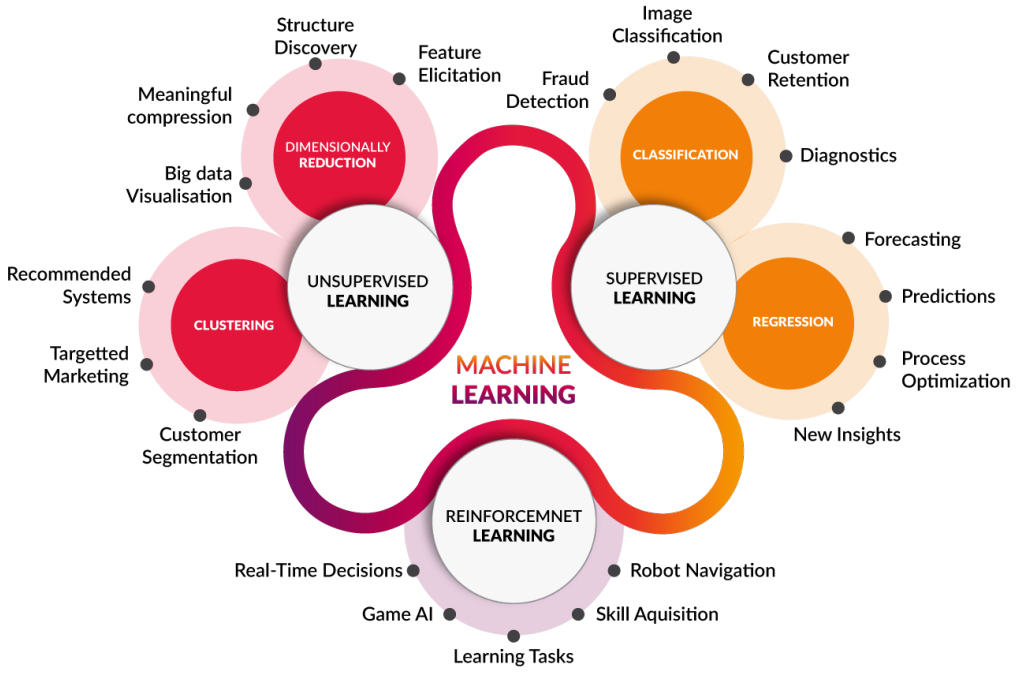



**[5] Izvor slike**:

`https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d`

# Tipovi mašinskog učenja

- Učenje potkrepljivanjem (reinforcement learning):
  - Donošenje odluka u realnom vremenu.
    - Autonomna vožnja vozila.
  - Navigacija robota.
  - Igranje igara.




**Izvor slike**:

`https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d`

# Šta su podaci?

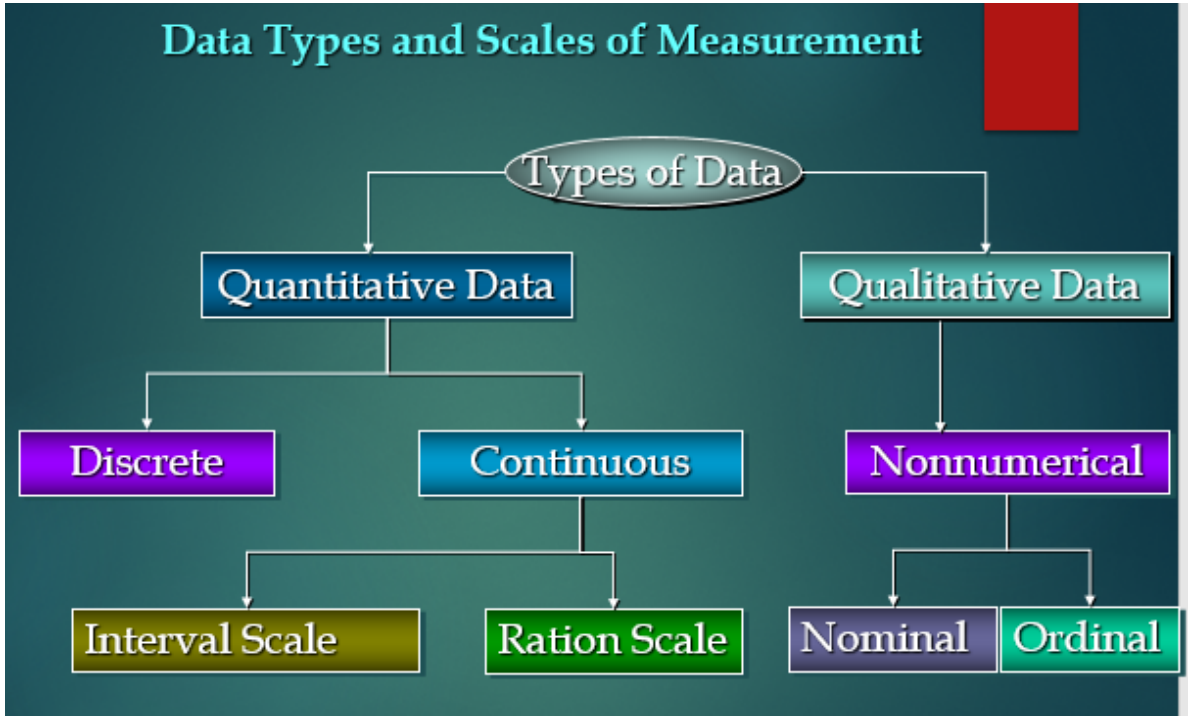
- Podaci su entiteti/pojmovi koje možemo da opišemo, izmerimo, analiziramo, skladištimo, generišemo.
- Podaci su kolekcija ili skup činjenica kao što su vrednosti ili opisi.

*Data refers to facts and statistics collected together for reference or analysis.*


**Izvor slike**: `googleimages`

# Podela

- Osnovna podela na osnovu pitanja- da li se podatak može opisati numerički:
  - Kvalitativni podaci,
  - Kvantitativni podaci.




**[5] Izvor slike**: `googleimages`

# Kvalitativni podaci

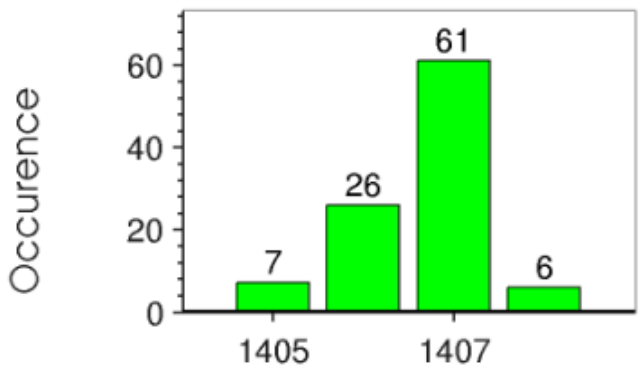
- Nisu numerički brojevi.
- Ne možemo ih izmeriti.
- Dozvoljavaju subjektivnost.
- Poznate i kao *kategoričke vrednosti*.
- Opisuju: ukus, boju, arhitektonski stil, bračno stanje, itd.

# Kvantitativni podaci

- Numerički brojevi.
- Možemo ih izmeriti.
- Odgovaraju na pitanja: *koliko puno* i *koliko često*.
- Nad njima se lako primenjuju statističke procedure.
- Razlikujemo dva tipa kvantitativnih podataka:
  1. Diskretni.
  2. Neprekidni.

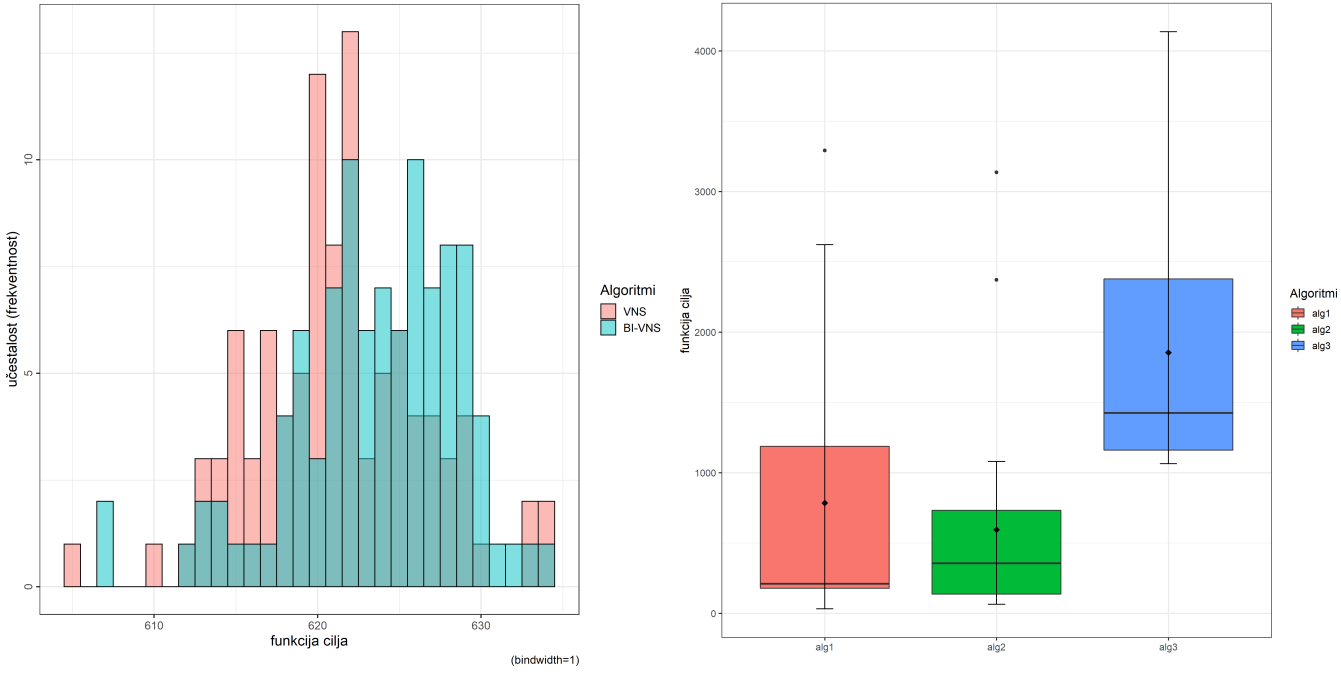
# Diskretni podaci

- Pokazuju prebrojavanja.
- Ne mogu se podeliti na razlomke ili decimalne vrednosti.
- Grafički prikaz su bar-chart.



# Neprekidni podaci

- Predstavljaju vrednosti koje se mogu podeliti na razlomke ili decimale.
- Neprekidne promenljive uzimaju vrednosti iz neograničenog skupa vrednosti.
- Primeri:
  - merenje vremena,
  - visina kod ljudi,
  - brzina automobila.



# Merne skale

- Podela mernih skala prema Stanley Stevens.
- Nominalna skala (nominal).
- Redna skala (ordinal).
- Intervalna skala (interval).
- Razmerna skala, skala odnosa (ratio).

# Nominalna skala

- Latinski *nomen*, eng. *name*.
- Zadovoljava osobinu identifikacije.
- Nominalni podaci su ponekada poznati kao *labels*.
- Ne postoji smislen radosled između entiteta.
- Ne možemo rangirati vrednosti na nominalnoj skali.
- Primer podataka na nominalnoj skali:
  - Rod: muški, ženski.
  - Boja kose: plava, braon.
  - Bračno stanje: oženjen/udata, sama(c). udovica/udovac.

# Redna skala

- Postoji smislen redosled između podataka.
- Ne možemo da računamo udaljenost između dva redna broja.
- Ne možemo primeniti aritmetičke operacije.
- Služe za rangiranje.
- Primer podataka na rednoj skali:
  - Ekonomski status: loš, srednji, dobar.
  - Rangiranje takmičara: prvi, drugi, treći.
  - Ocene: odličan, dobar, loš.

# Intervalna skala

- Postoji smislen redosled između podataka tj postoji uređenje.
- Možemo da računamo udaljenost između dve vrednosti.
- Možemo primeniti aritmetičke operacije, bez množenja i deljenja.
- Može da sadrži negativne brojeve.
- Primer podataka na intervalnoj skali:
  - Temperatura.
  - Datumi.
  - Rezulati testa inteligencije.

# Razmerna skala

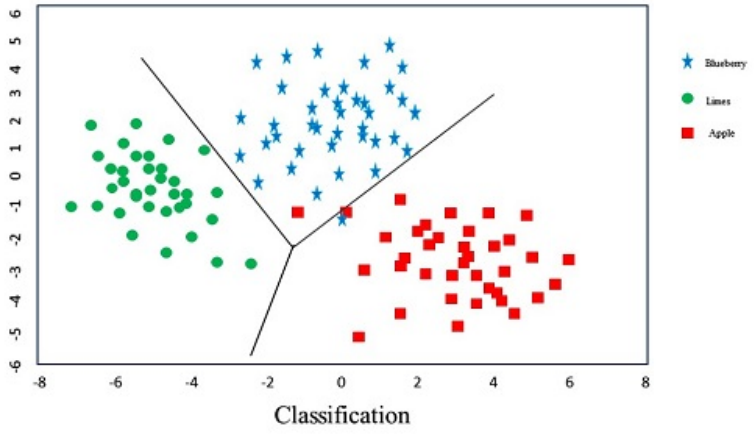
- Obuhvata neke osobine intervalne skale, ali sada postoji asplutno nulta vrednost (odsustvo nečeg).
- Zadovoljava osobine identifikacije, rasporeda, veličine i razlike.
- Ne sadrži negativne brojeve.
- Možemo primeniti sve aritmetičke operacije.
- Primer podataka na razmernoj skali:
  - Visina
  - Težina
  - Dužina

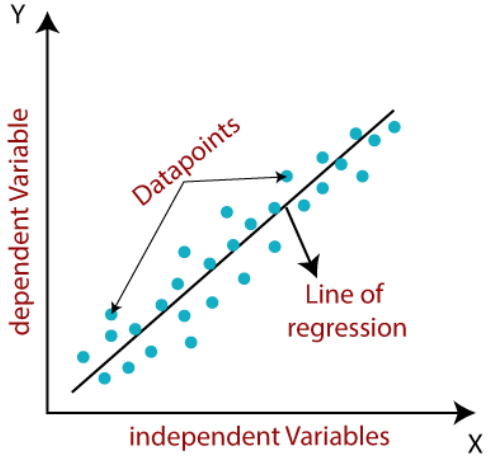
# Ključna terminologija

- Oznaka (label)
- Atribut (feature)
  - Za elektronsku poštu: *naslov*, *dužina*, *prva reč*,...
  - Za nekretninu: *kvadratura*, *lokacija*, *broj soba*,...

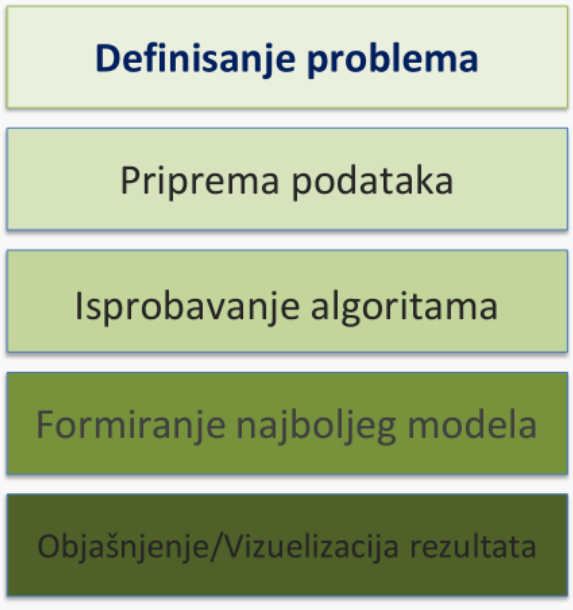
# Zadaci nadgledanog mašinskog učenja

- Klasifikacija - izlaz je jedna vrednost iz konačnog skupa vrednosti.
- Razlikujemo:
  - Binarnu klasifikaciju (npr. "položio ispit", "nije položio ispit").
  - Višeklasna klasifikacija (npr. za dato voće da li je "jabuka", "kruška" ili "pomorandža").
- Regresija - izlaz je neprekidna vrednost.



# Nadgledano učenje




1. **Definisanje problema**

**Šta je problem?**
**Šta je cilj rešavanja problema?**
**Šta se dobija rešavanjem problema?**
**Kako bismo rešili problem?**

**Izvor**: `https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474`

# Nadgledano učenje

Neformalni opis problema: *Potreban nam je program koji će moći da predviđa cenu stana na osnovu kvadrature i lokacije stana.*

*T – Određivanje cene stana*

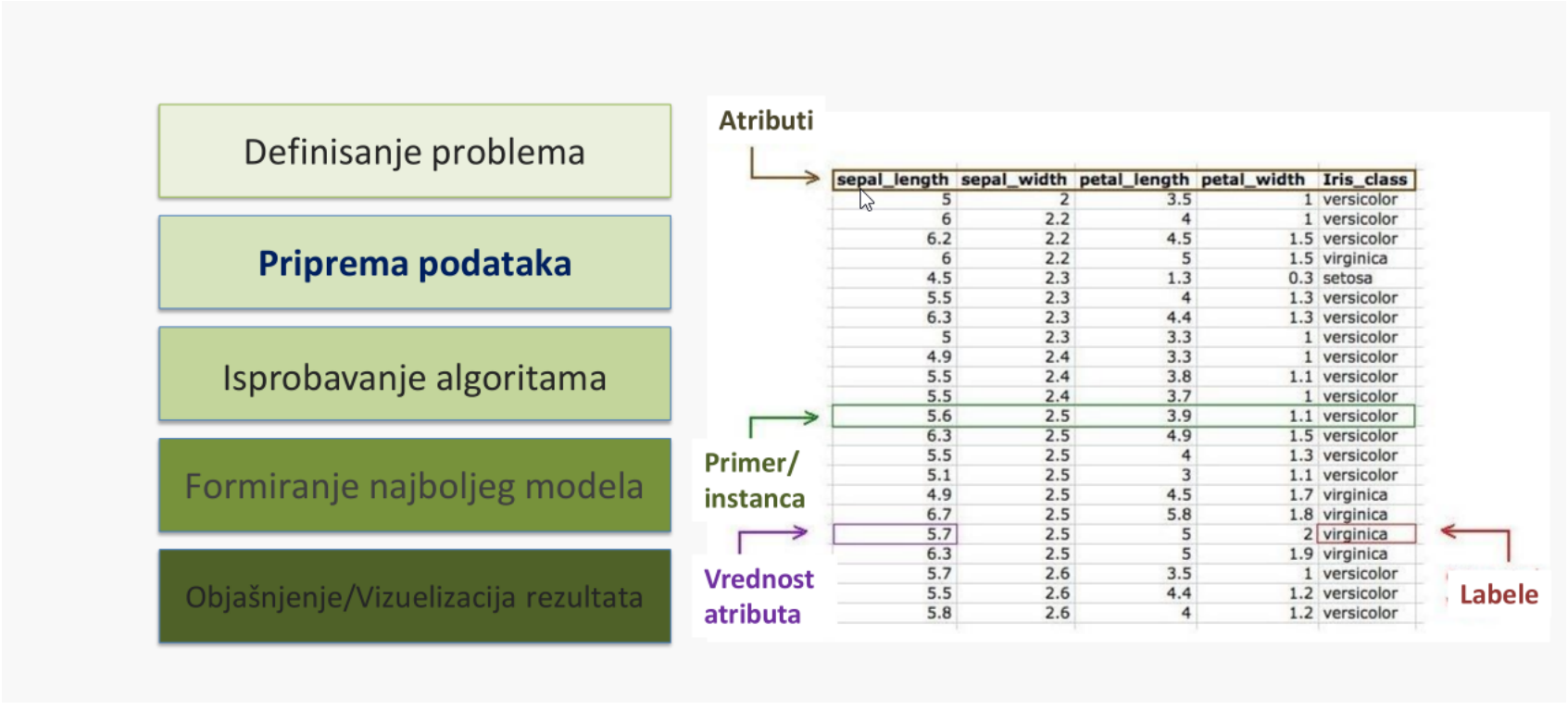
*E - Podaci o kvadraturama, lokacijama i cenama stanova*

*P – razlika predviđene i stvarne cene za stanove koji nisu bili deo podataka na kojima je program sticao iskustvo.*



**Izvor**: https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474

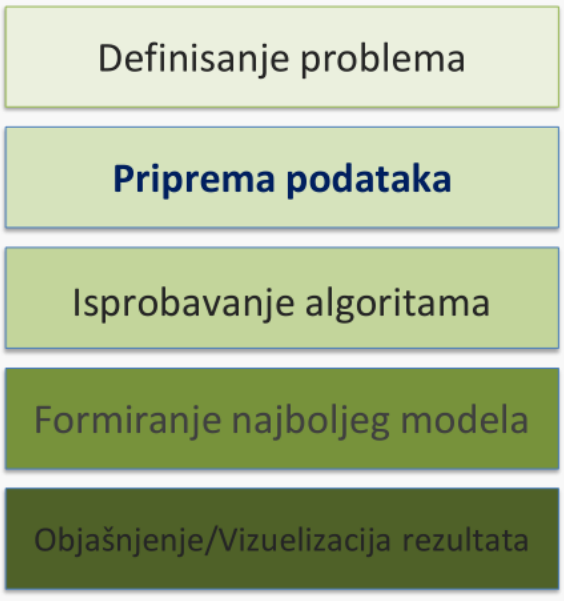
# Nadgledano učenje




| sepal_length | sepal_width | petal_length | petal_width | Iris_class |
|---|---|---|---|---|
| 5 | 2 | 3.5 | 1 | versicolor |
| 6 | 2.2 | 4 | 1 | versicolor |
| 6.2 | 2.2 | 4.5 | 1.5 | versicolor |
| 6 | 2.2 | 5 | 1.5 | virginica |
| 4.5 | 2.3 | 1.3 | 0.3 | setosa |
| 5.5 | 2.3 | 4 | 1.3 | versicolor |
| 6.3 | 2.3 | 4.4 | 1.3 | versicolor |
| 5 | 2.3 | 3.3 | 1 | versicolor |
| 4.9 | 2.4 | 3.3 | 1 | versicolor |
| 5.5 | 2.4 | 3.8 | 1.1 | versicolor |
| 5.5 | 2.4 | 3.7 | 1 | versicolor |
| 5.6 | 2.5 | 3.9 | 1.1 | versicolor |
| 6.3 | 2.5 | 4.9 | 1.5 | versicolor |
| 5.5 | 2.5 | 4 | 1.3 | versicolor |
| 5.1 | 2.5 | 3 | 1.1 | versicolor |
| 4.9 | 2.5 | 4.5 | 1.7 | virginica |
| 6.7 | 2.5 | 5.8 | 1.8 | virginica |
| 5.7 | 2.5 | 5 | 2 | virginica |
| 6.3 | 2.5 | 5 | 1.9 | virginica |
| 5.7 | 2.6 | 3.5 | 1 | versicolor |
| 5.5 | 2.6 | 4.4 | 1.2 | versicolor |
| 5.8 | 2.6 | 4 | 1.2 | versicolor |

**Izvor**: `https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474`

# Nadgledano učenje




1. **Priprema podataka**

**Analiza**
**Odabir**
**Pretprocesiranje**
**Transformisanje**

- Pregled i vizuelizacija atributa
- Odnosi između atributa.
- Razmišljanje o podacima u kontekstu problema.

**Izvor**: `https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474`

# Nadgledano učenje - podaci

*heart_disease* skup podataka sadrži sledeće atribute:

- **age**: continuous
- **sex**: categorical, 2 values {0: female, 1: male}
- **cp** (chest pain type): categorical, 4 values {1: typical angina, 2: atypical angina, 3: non-angina, 4: asymptomatic angina}
- **restbp** (resting blood pressure on admission to hospital): continuous (mmHg)
- **chol (serum cholesterol level)**: continuous (mg/dl)
- **fbs** (fasting blood sugar): categorical, 2 values {0: <= 120 mg/dl, 1: > 120 mg/dl}
- **restecg** (resting electrocardiography): categorical, 3 values {0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy}
- **thalach** (maximum heart rate achieved): continuous
- **exang** (exercise induced angina): categorical, 2 values {0: no, 1: yes}
- **oldpeak** (ST depression induced by exercise relative to rest): continuous
- **slope** (slope of peak exercise ST segment): categorical, 3 values {1: upsloping, 2: flat, 3: downsloping}
- **ca** (number of major vessels colored by fluoroscopy): discrete (0,1,2,3)
- **thal**: categorical, 3 values {3: normal, 6: fixed defect, 7: reversible defect}
- **num** (diagnosis of heart disease): categorical, 5 values {0: less than 50% narrowing in any major vessel, 1-4: more than 50% narrowing in 1-4 vessels}

**Izvor**: `https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474`

# Nadgledano učenje - transformacija

```
# load the dataset
heart_df = pd.read_csv('../data/heart_disease.csv', header=None, names=columns)
heart_df.head()
```

| | age | sex | cp | restbp | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | num |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63.0 | 1.0 | 1.0 | 145.0 | 233.0 | 1.0 | 2.0 | 150.0 | 0.0 | 2.3 | 3.0 | 0.0 | 6.0 | 0.0 |
| 1 | 67.0 | 1.0 | 4.0 | 160.0 | 286.0 | 0.0 | 2.0 | 108.0 | 1.0 | 1.5 | 2.0 | 3.0 | 3.0 | 2.0 |
| 2 | 67.0 | 1.0 | 4.0 | 120.0 | 229.0 | 0.0 | 2.0 | 129.0 | 1.0 | 2.6 | 2.0 | 2.0 | 7.0 | 1.0 |
| 3 | 37.0 | 1.0 | 3.0 | 130.0 | 250.0 | 0.0 | 0.0 | 187.0 | 0.0 | 3.5 | 3.0 | 0.0 | 3.0 | 0.0 |
| 4 | 41.0 | 0.0 | 2.0 | 130.0 | 204.0 | 0.0 | 2.0 | 172.0 | 0.0 | 1.4 | 1.0 | 0.0 | 3.0 | 0.0 |

**Izvor**: https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474

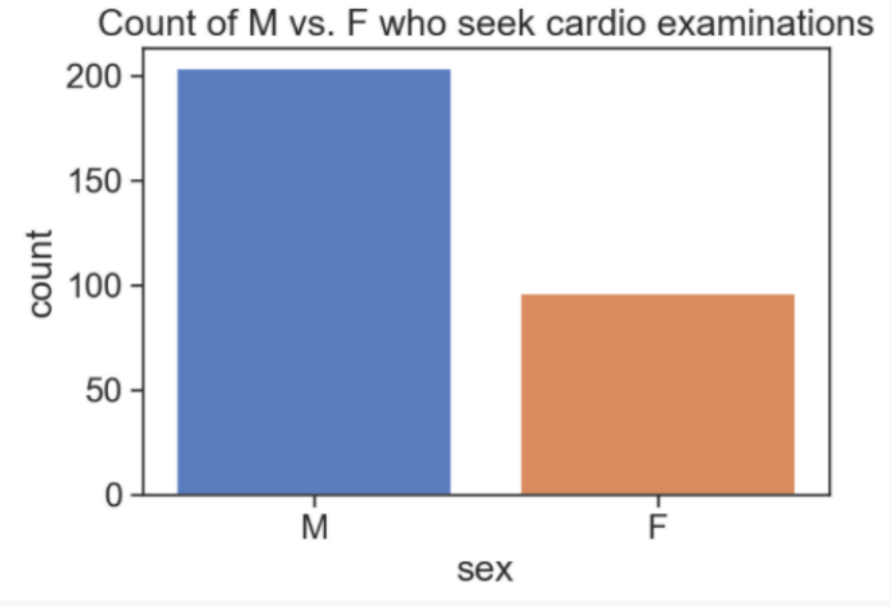
# Nadgledano učenje - analiza podataka

1.At what ages do people seek cardiological exams?


**Izvor**: https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474

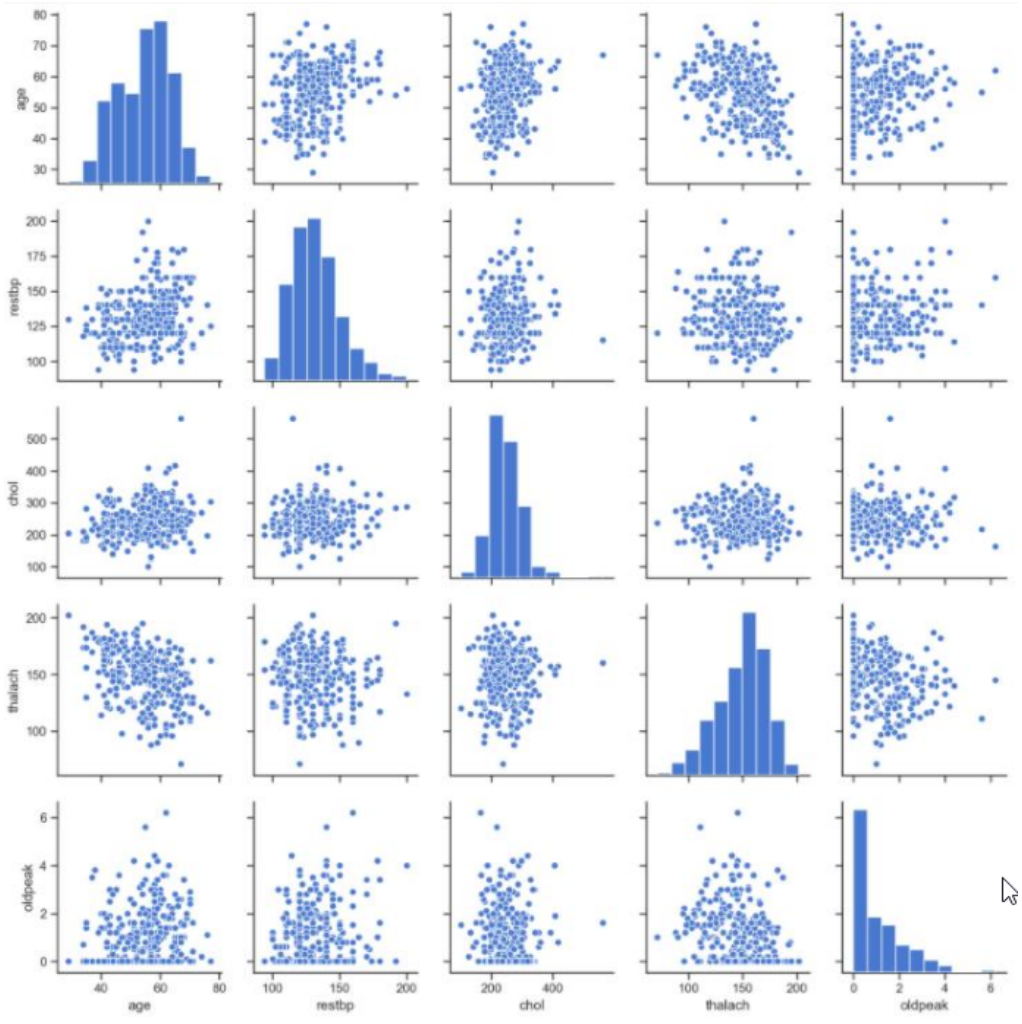
# Nadgledano učenje - analiza podataka

2. Do men seek help more than women?




**Izvor**: https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474

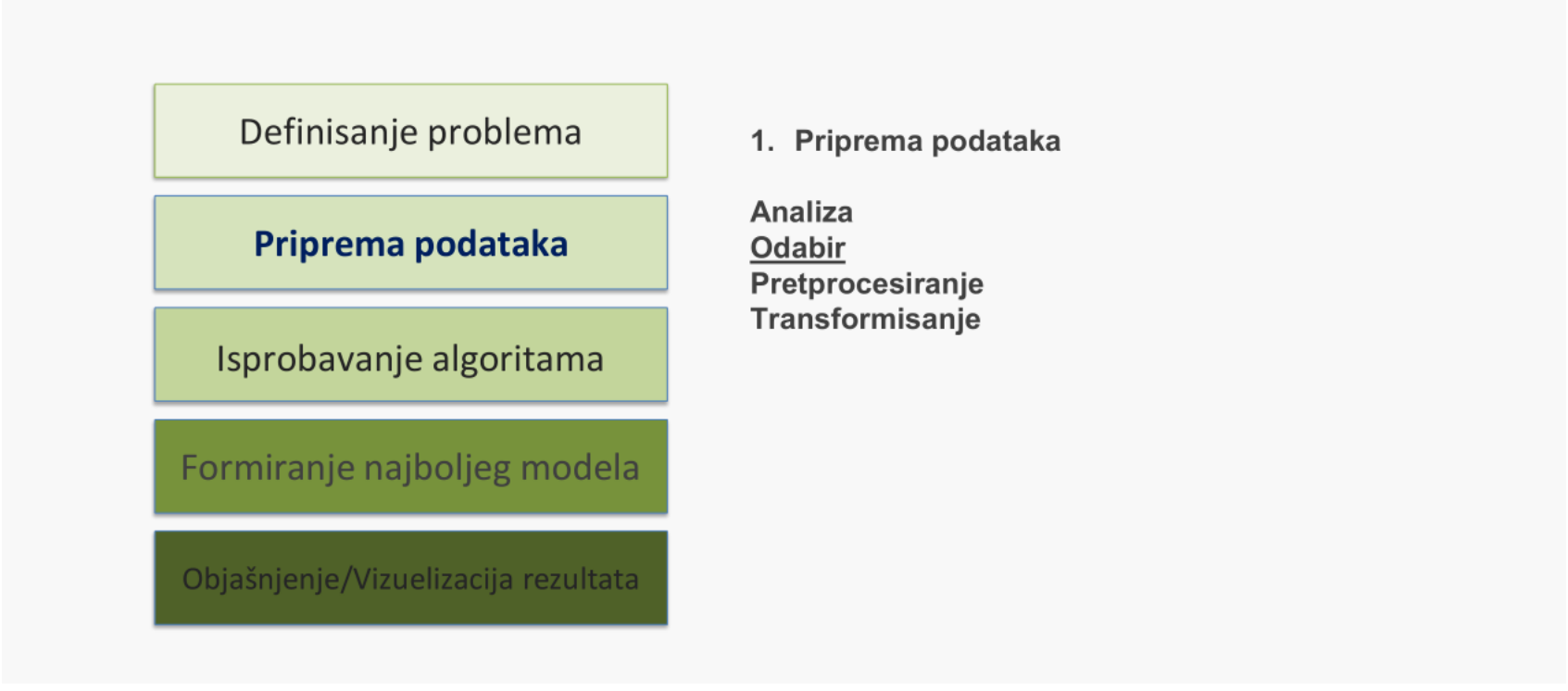
# Nadgledano učenje - analiza podataka




**Izvor**:

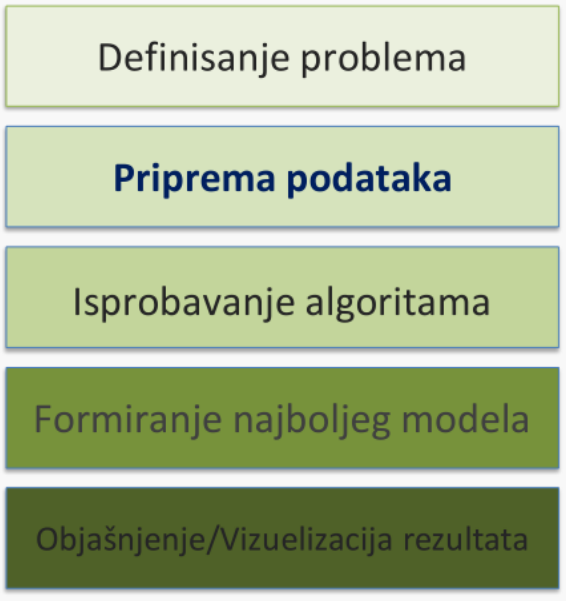
https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474

# Nadgledano učenje - odabir




**Izvor**: `https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474`

# Nadgledano učenje - odabir



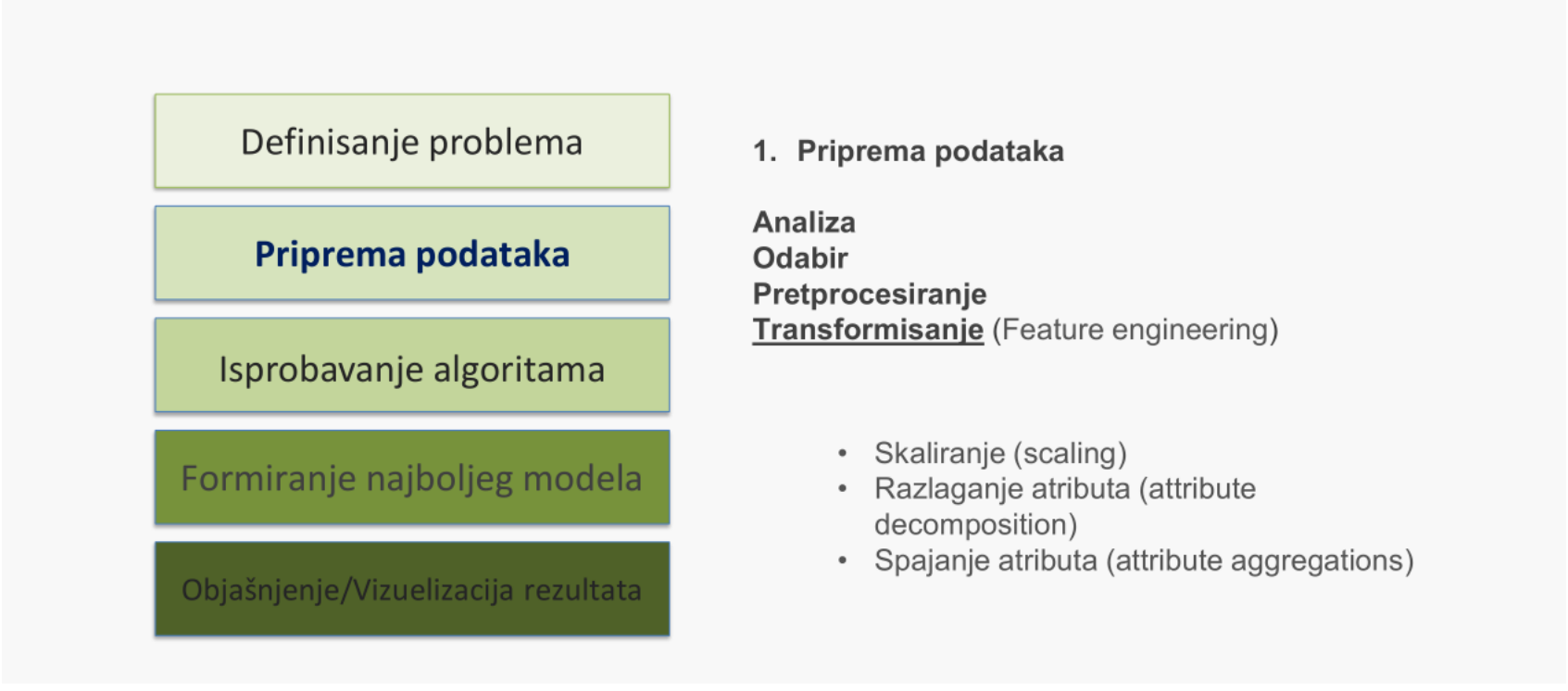

**1. Priprema podataka**

**Analiza**
**Odabir**
**Pretprocesiranje**
**Transformisanje**

- Formatiranje (formatting),
- Prečišćavanje (cleaning)
- Uzorkovanje (sampling)

**Izvor**: `https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474`

# Nadgledano učenje - odabir




**Izvor**: `https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474`

# Nadgledano učenje - izbor

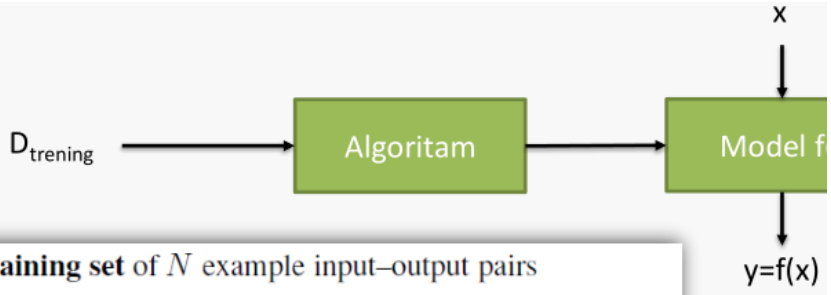
- Izbor modela za učenje sa raznim pretpostavkama: npr koji parametri tog modela su dobri.
- Neki modeli više odgovaraju za rad sa slikama, druge više za rad sa tekstom ili audio zapisom, treći bolje rade sa numeričkim podacima.
- Algoritmi za klasifikaciju ili za regresiju.

# Nadgledano učenje - obuka

Cilj algoritama mašinskog učenja je određivanje funkcije $f: X \to Y$ koja preslikava **ulaz** $X$ u **izlaz** $Y$.

$f$ – model, hipoteza, prediktor

Algoritam uči funkciju $f$ iz trening podataka.

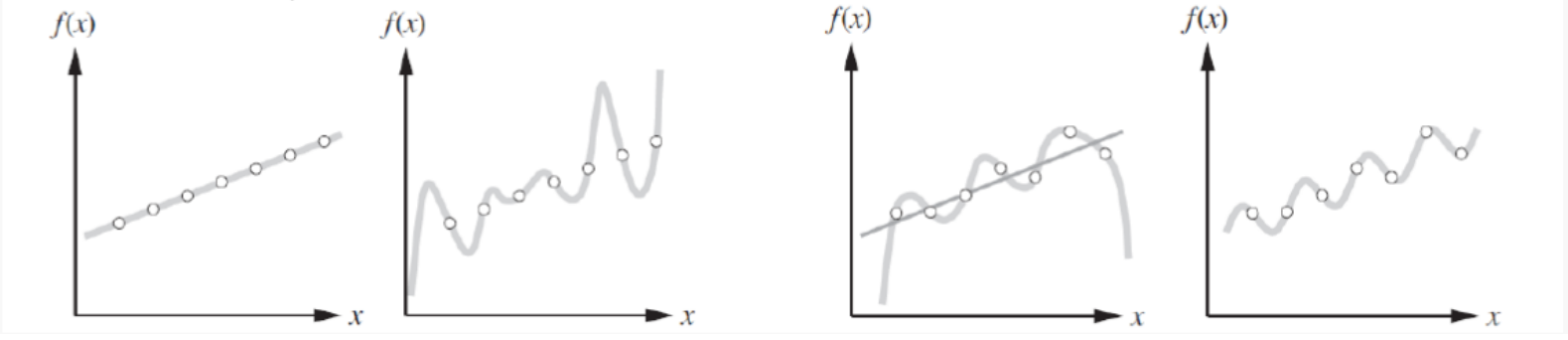



Given a **training set** of $N$ example input–output pairs

$$(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N) ,$$

where each $y_j$ was generated by an unknown function $y = f(x)$, discover a function $h$ that approximates the true function $f$.

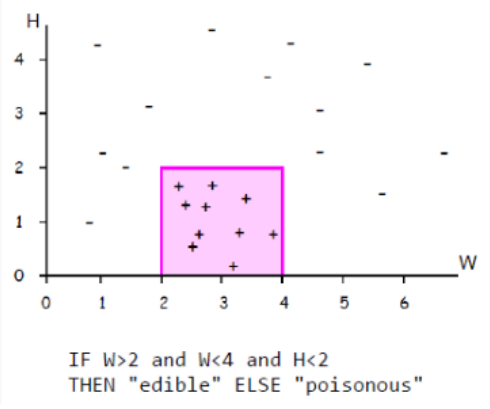
Primeri različitih modela $f$ za jedan trening skup:

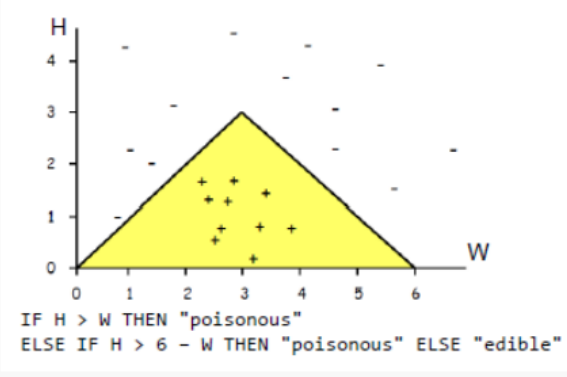


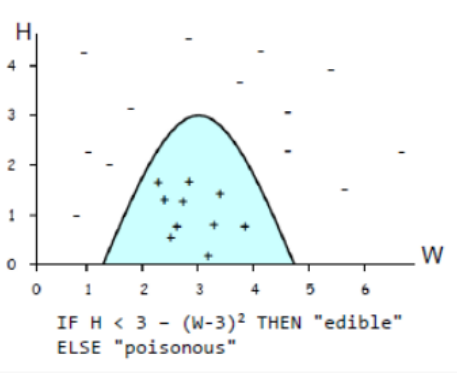

**Izvor**: https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474

# Nadgledano učenje - evaluacija

Rezulat učenja je model - klasifikator koji je u stanju da klasifikuje nove pečurke
Klasifikator može biti dat i u formi if-then pravila:




**Izvor**: https://imi.pmf.kg.ac.rs/moodle/course/view.php?id=474

# Naredna predavanja

- Generalizacija i preprilagođavanje, Okamova oštrica.
- Probabilistički modeli (linearna i logistička regresija, naivni Bajes).

Hvala na pažnji.

Da li imate pitanja?