MAS Nauka o podacima

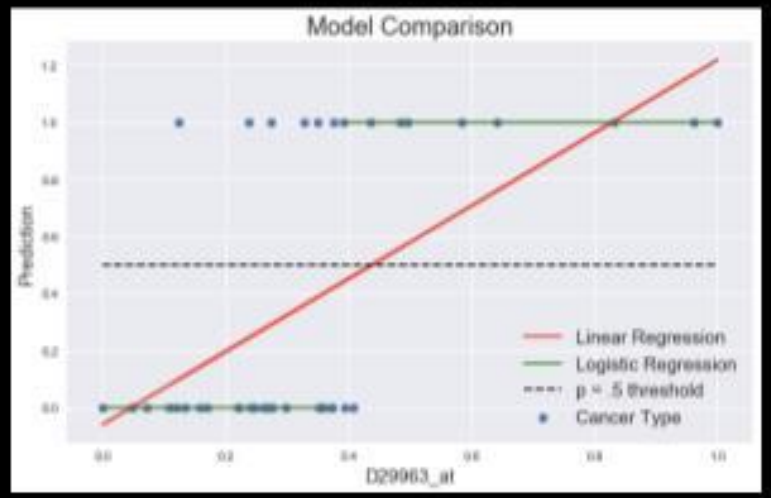
MAŠINSKO UČENJE 1

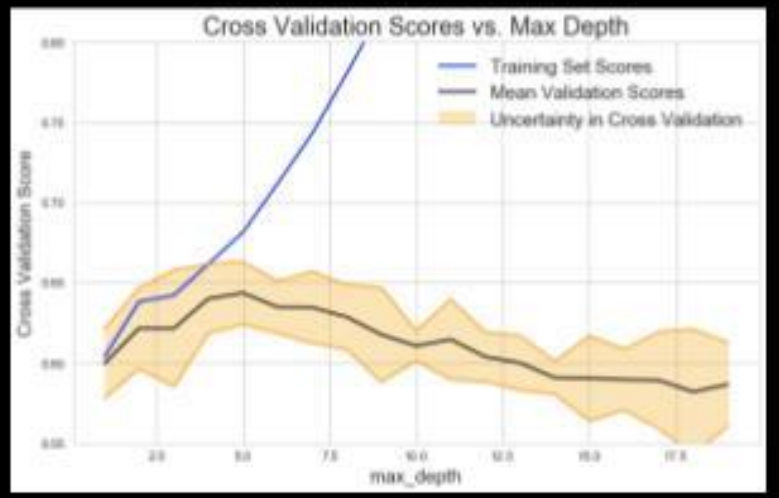
what my friends think I do

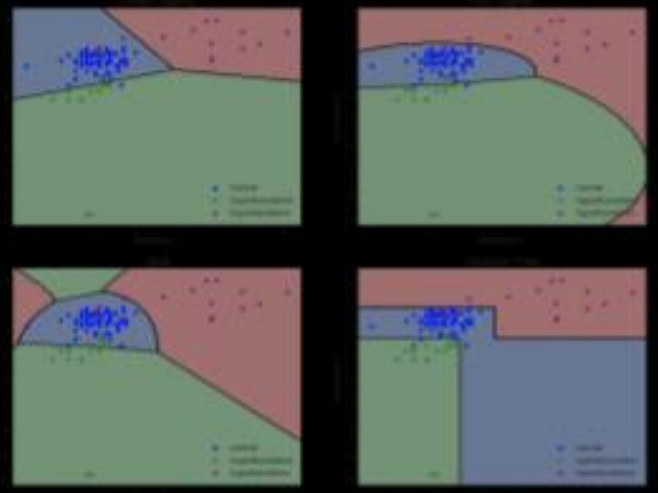
what my family thinks I do

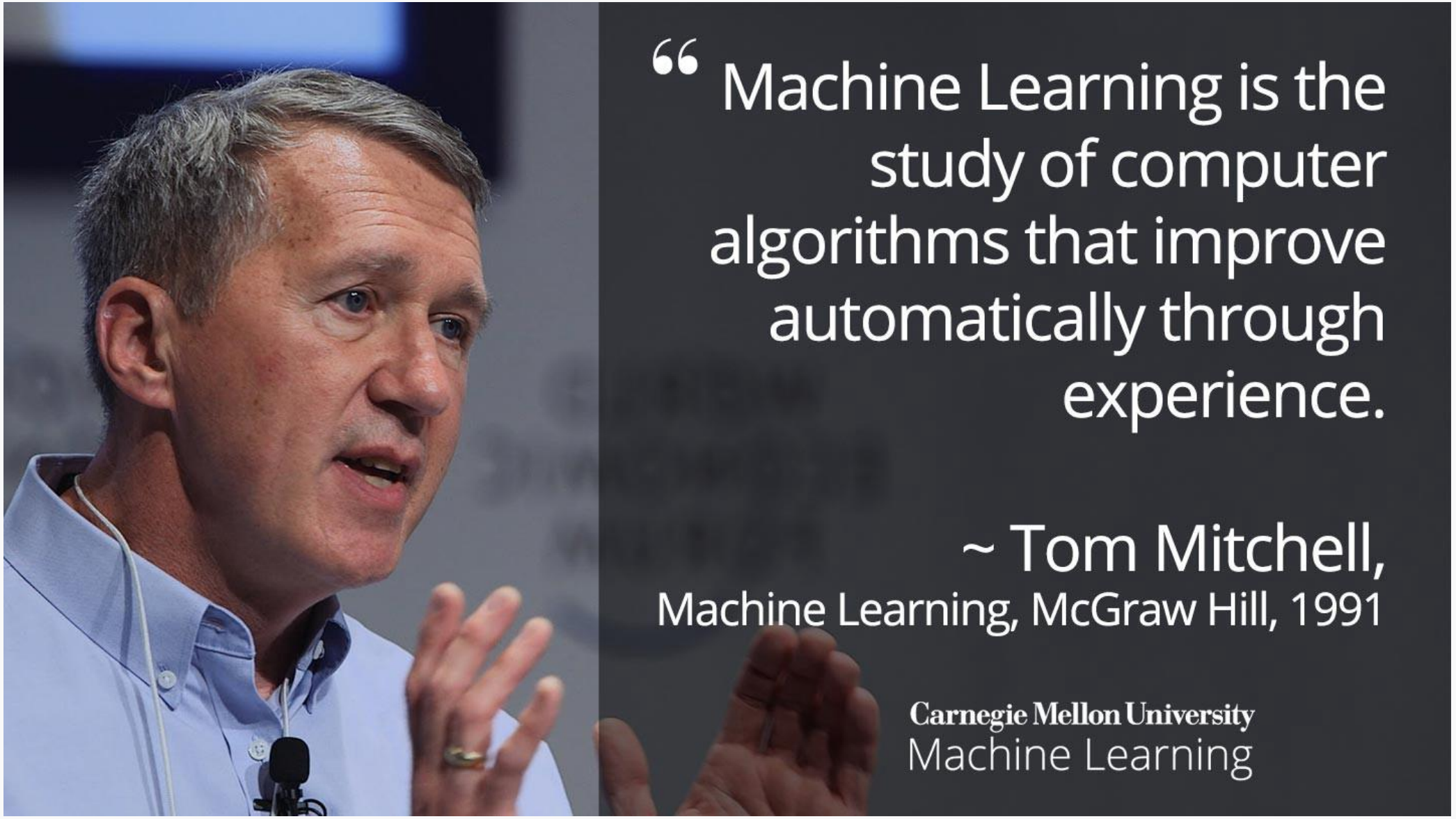
what society thinks I do

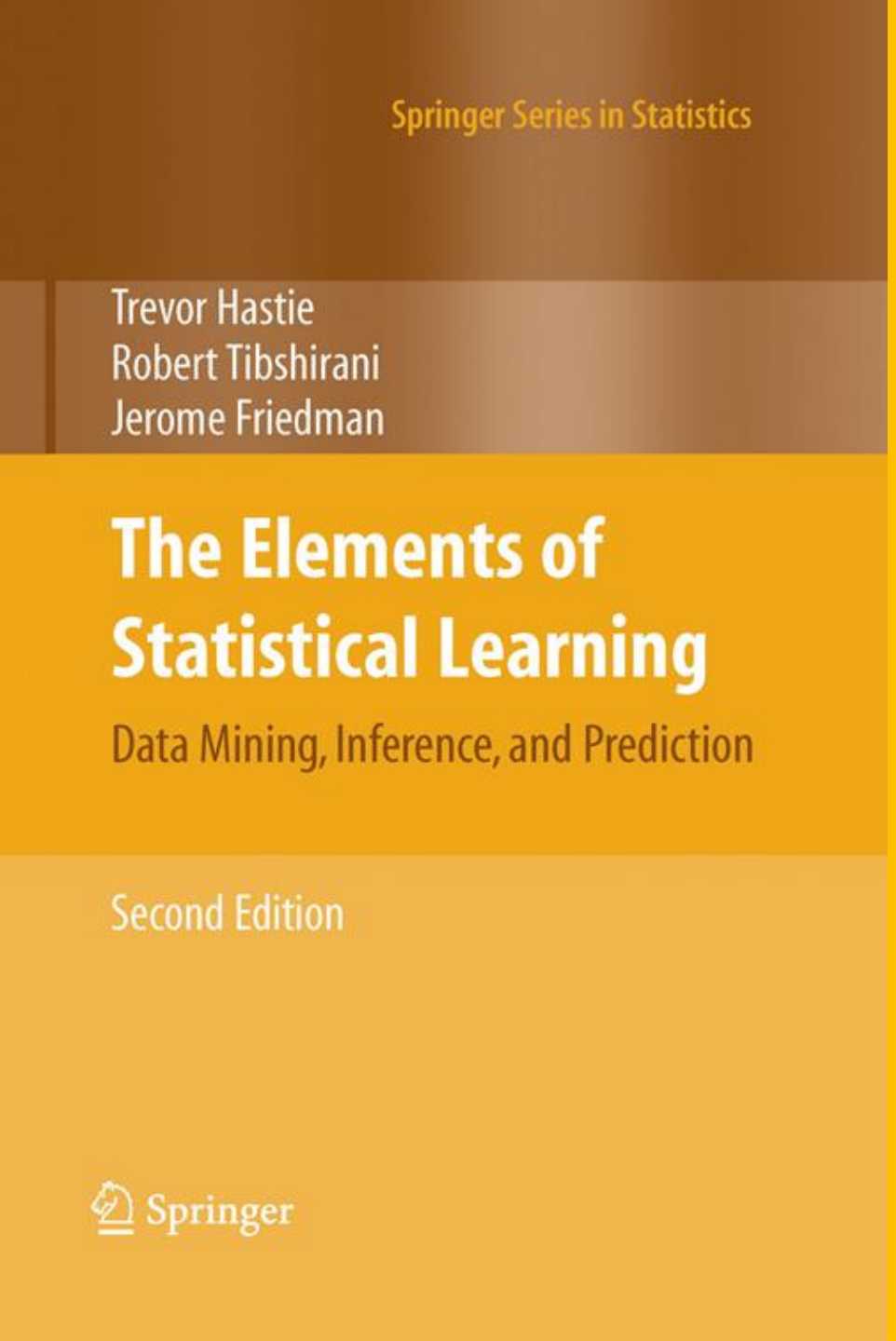
what I actually (will) do in Data Science







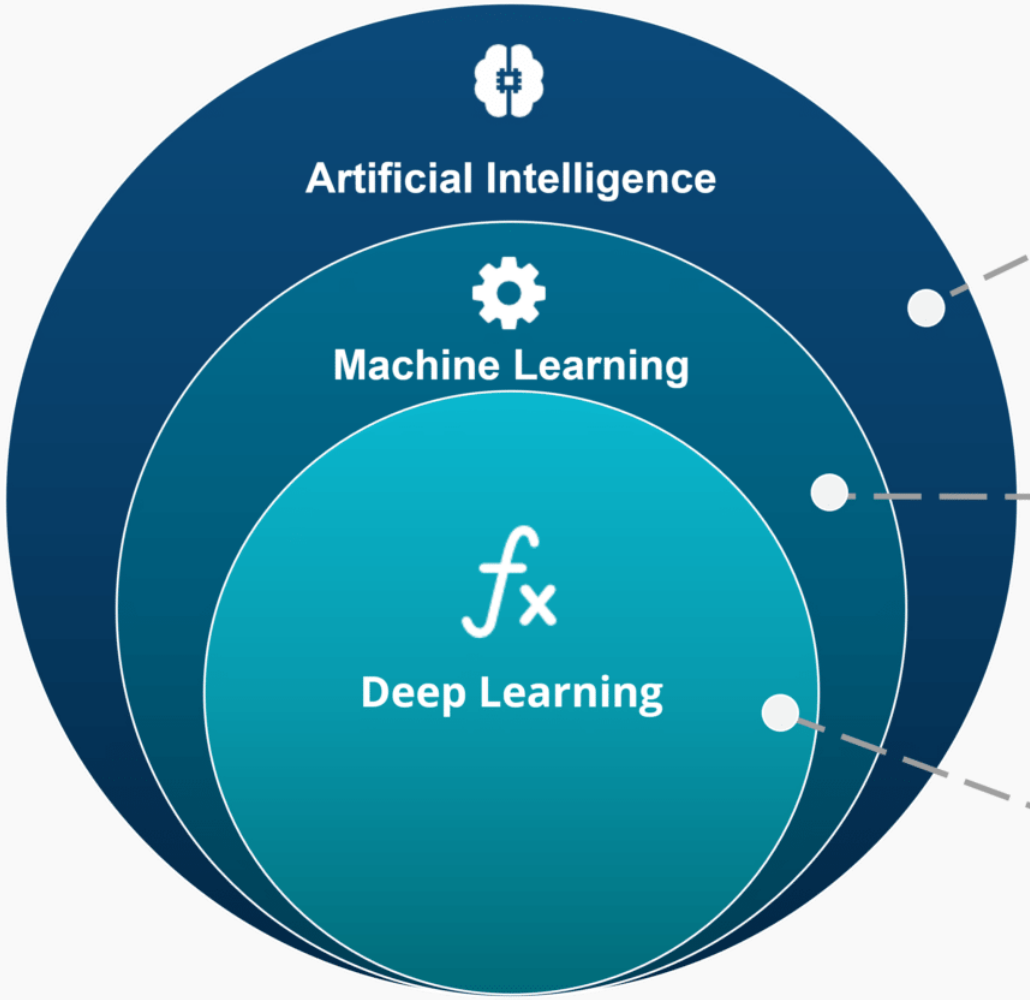

“ Machine Learning is the study of computer algorithms that improve automatically through experience.
~ Tom Mitchell,
Machine Learning, McGraw Hill, 1991
Carnegie Mellon University
Machine Learning



„Vast amounts of data are being generated in many fields, and the statisticians's job is to make sense of it all: to extract important patterns and trends, and to understand "what the data says". We call this learning from data."

AI vs ML

Mašinsko učenje je podoblast Veštačke inteligencije



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

Vrste mašinskog učenja

Nadgledano učenje (*supervised learning*)

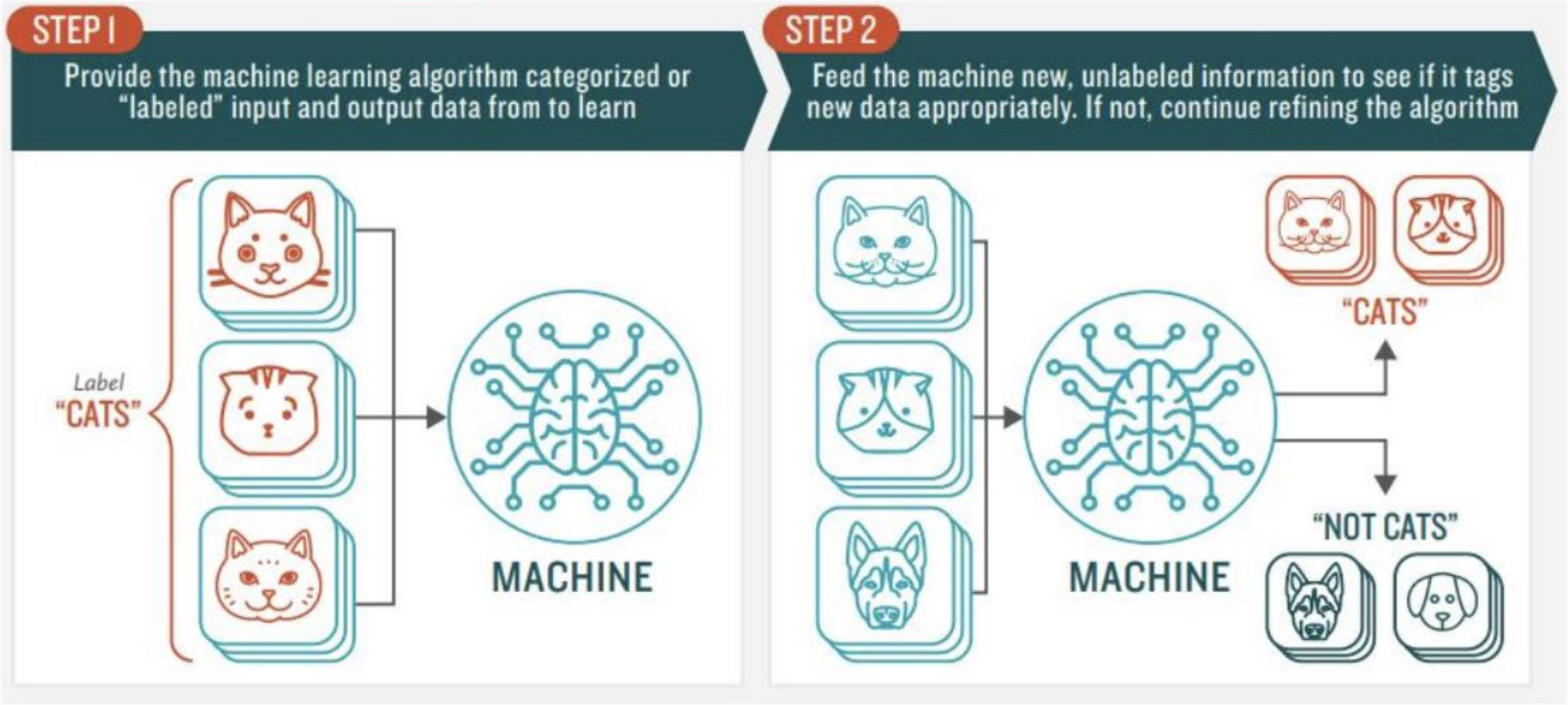
Nenadgledano učenje (*unsupervised learning*)

Učenje sa pojačavanjem (*reinforcement learning*)

Nadgledano obučavanje

Primeri za učenje dati su u obliku parova vrednosti ulaza i izlaza.

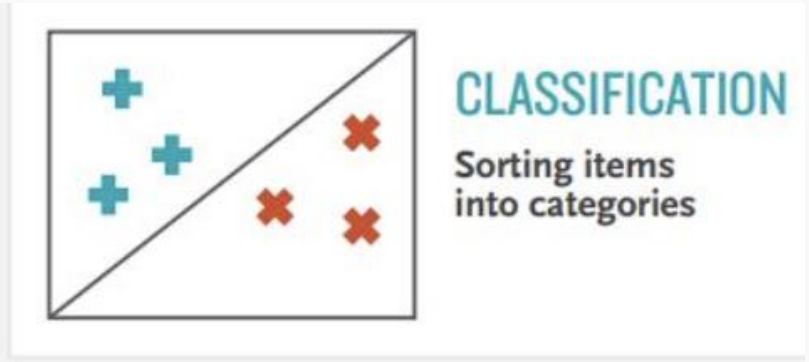
Uči se funkcija koja preslikava ulaze u izlaze.



Zadaci nadgledanog mašinskog učenja

Klasifikacija - izlaz je jedna vrednost iz konačnog skupa vrednosti

- Binarna klasifikacija (jedna od dve moguće klase, npr. za dati mejl odrediti da li „jeste spam“ ili „nije spam“)
- Višeklasna klasifikacija (jedna od konačno mnogo mogućih klasa, npr. za dato voće odrediti da li je pomorandža, grejpfrut ili mandarina)

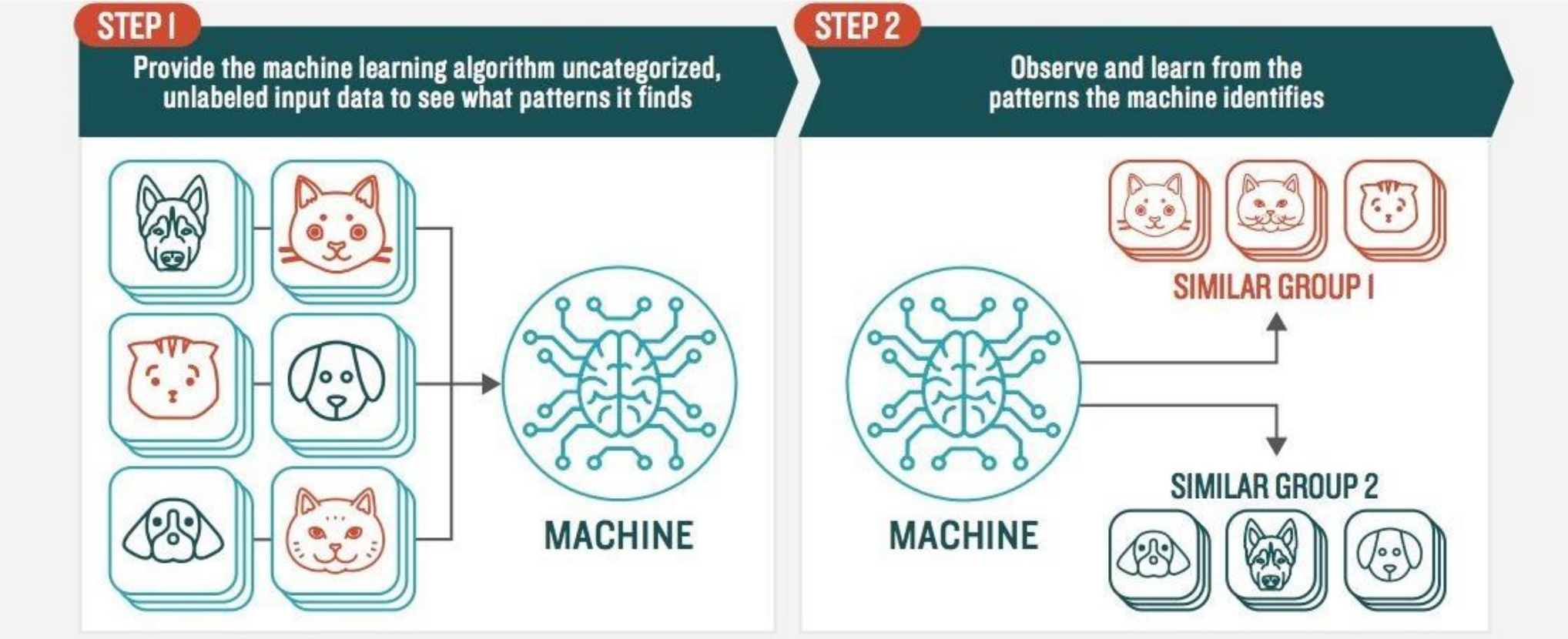


Regresija - izlaz je kontinualna vrednost (npr. predvideti cenu kuće na osnovu njene lokacije i veličine)

Binarna klasifikacija može biti podvedena pod regresioni problem sa labelama -1 i +1.

Nenadgledano obučavanje

Donosi zaključke i pronalazi obrasce iz ulaznih podataka koji nisu označeni.



Zadaci nenadgledanog mašinskog učenja

Klasterizacija (Clustering)

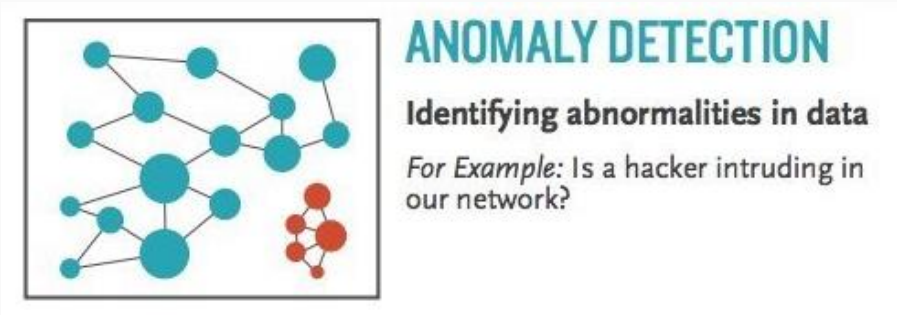
Prepoznavanje anomalija u podacima (Anomaly detection)

Smanjenje dimenzija (Dimensionality reduction).

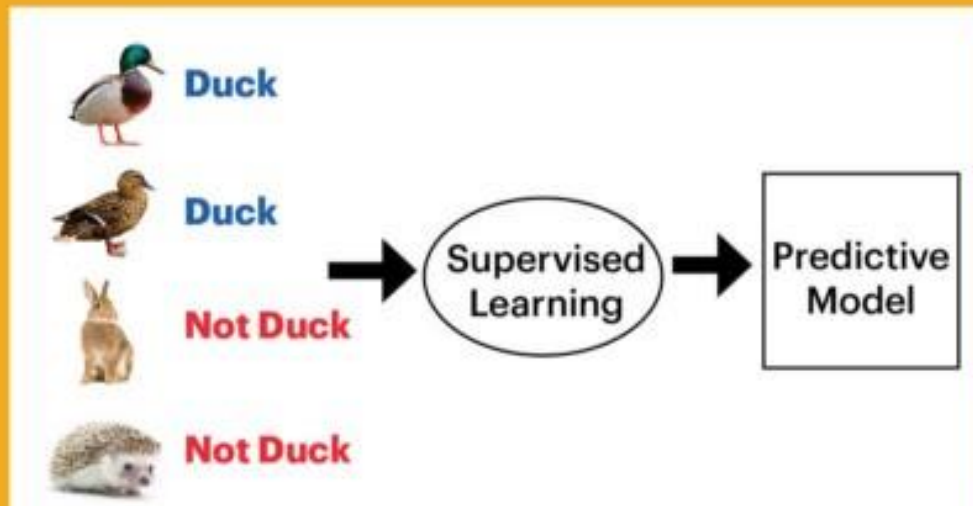
Grupisanje je otkrivanje potencijalno korisnih grupa ulaznih primera koji su po nečemu slični (segmentacija klijenata). Uobičajene tehnike klasterizacije su k-means, hierarchical clustering, mean shift clustering, i density-based clustering.

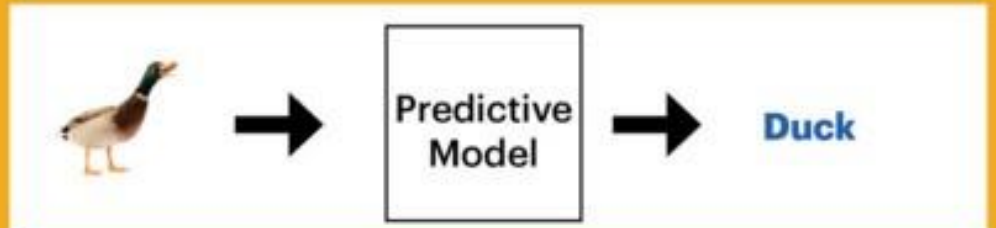
Smanjenje dimenzija je proces redukovanja broj atributa i izdvajanje onih koji imaju najveći značaj za problem koji se rešava. Uobičajena tehnika je PCA (Principal Component Analysis)

Otkrivanje anomalija je identifikacija retkih pojava (događaja) koje izazivaju sumnje značajnim razlikovanjem od većine podataka. Anomalne aktivnosti mogu biti povezane sa nekom vrstom problema ili retkim događajima kao što su bankarske prevare, medicinski problemi, strukturni nedostaci, neispravnost opreme itd.

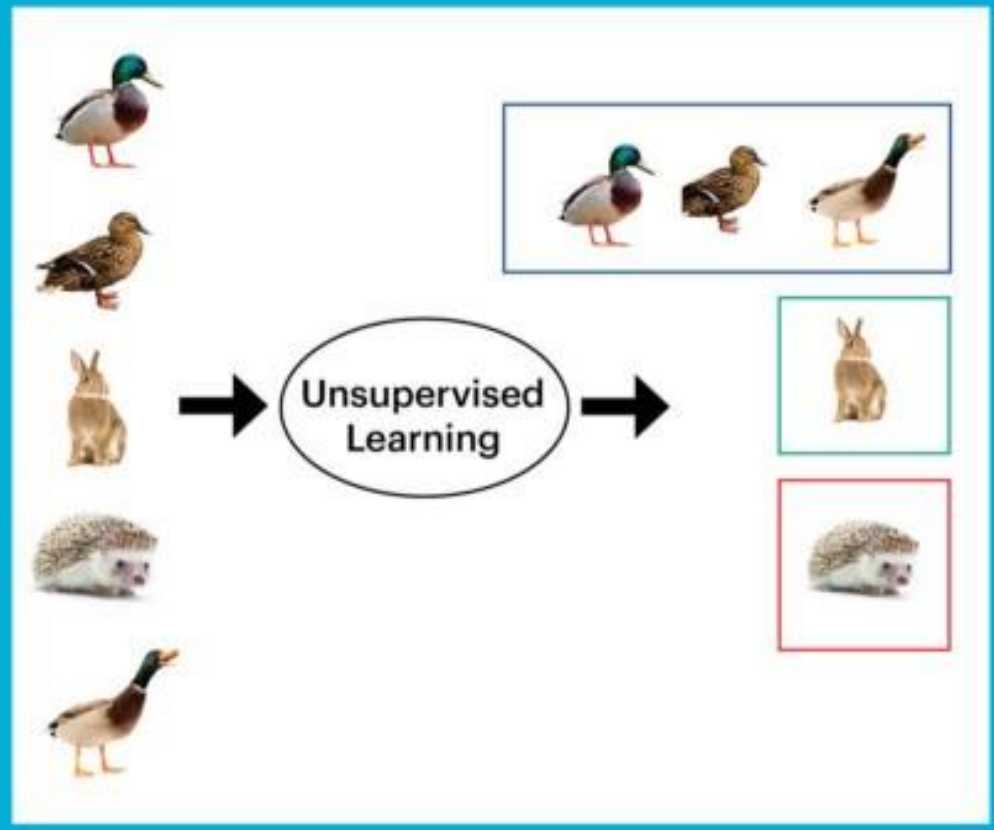


Supervised Learning (Classification Algorithm)





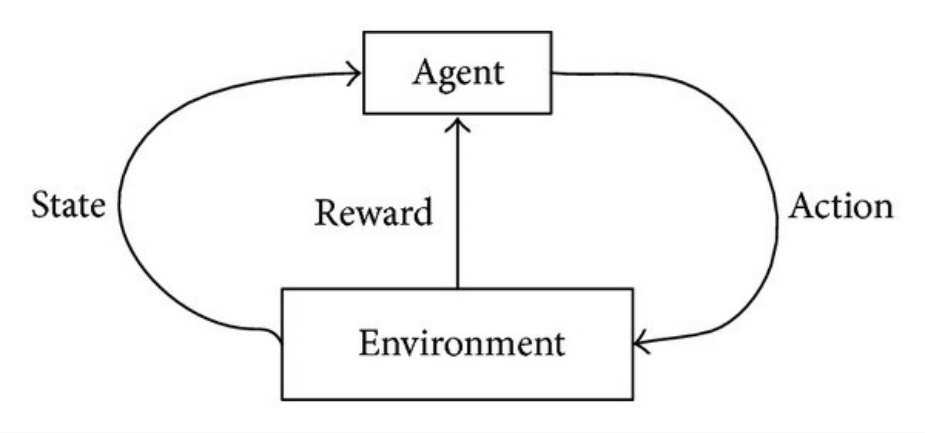
Unsupervised Learning (Clustering Algorithm)

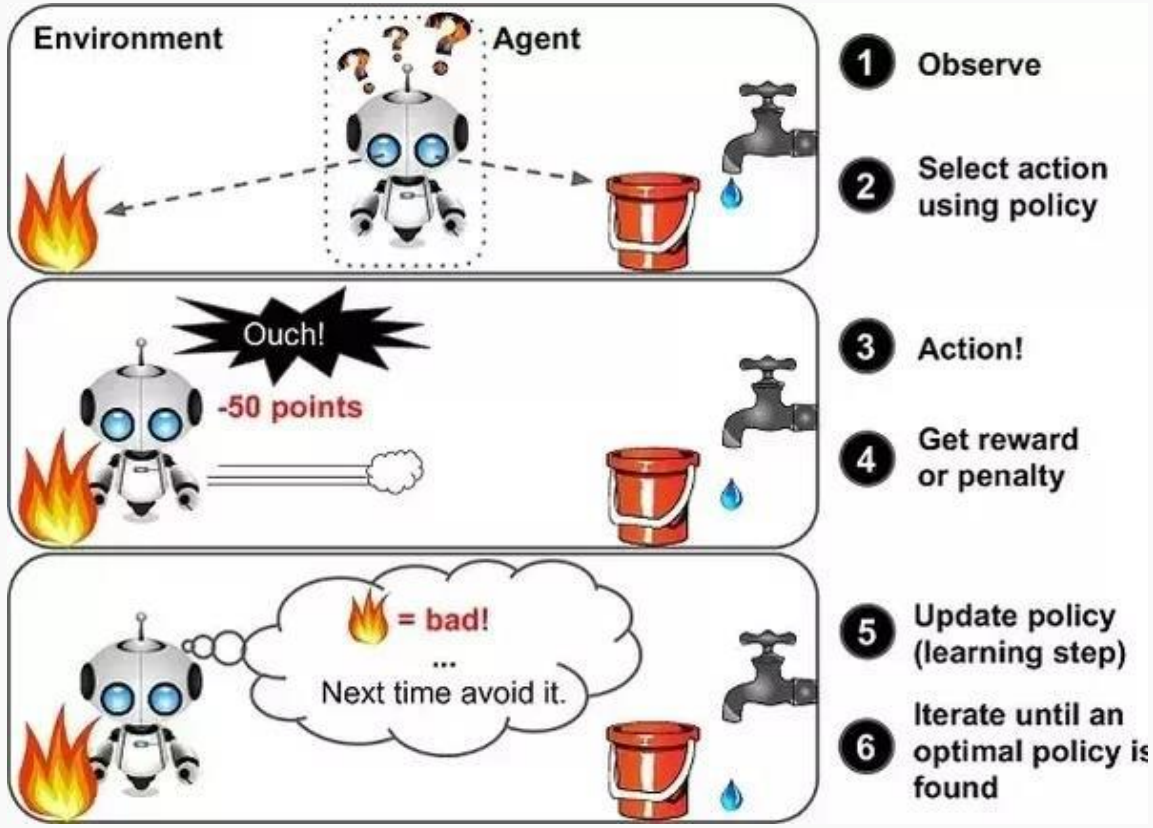


Učenje sa pojačavanjem (Reinforcement learning)

Agent uči iz niza pojačavanja – nagrada ili kazni.

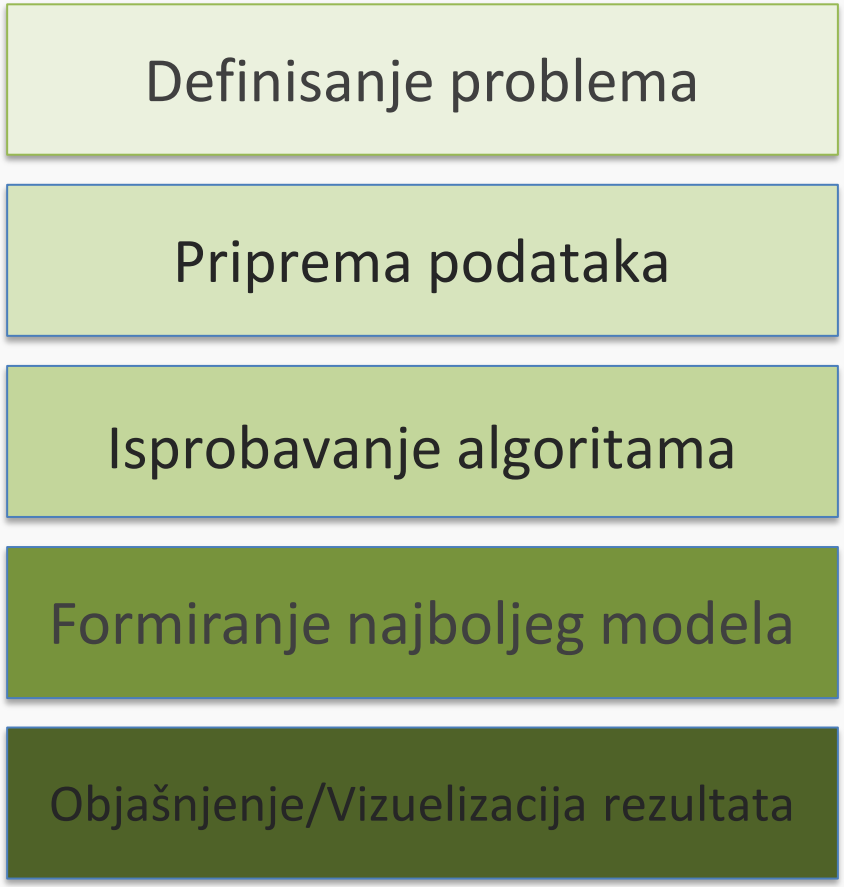
Na primer, nedostatak napojnice na kraju vožnje daje taksi agentu indikaciju da nešto nije bilo u redu.



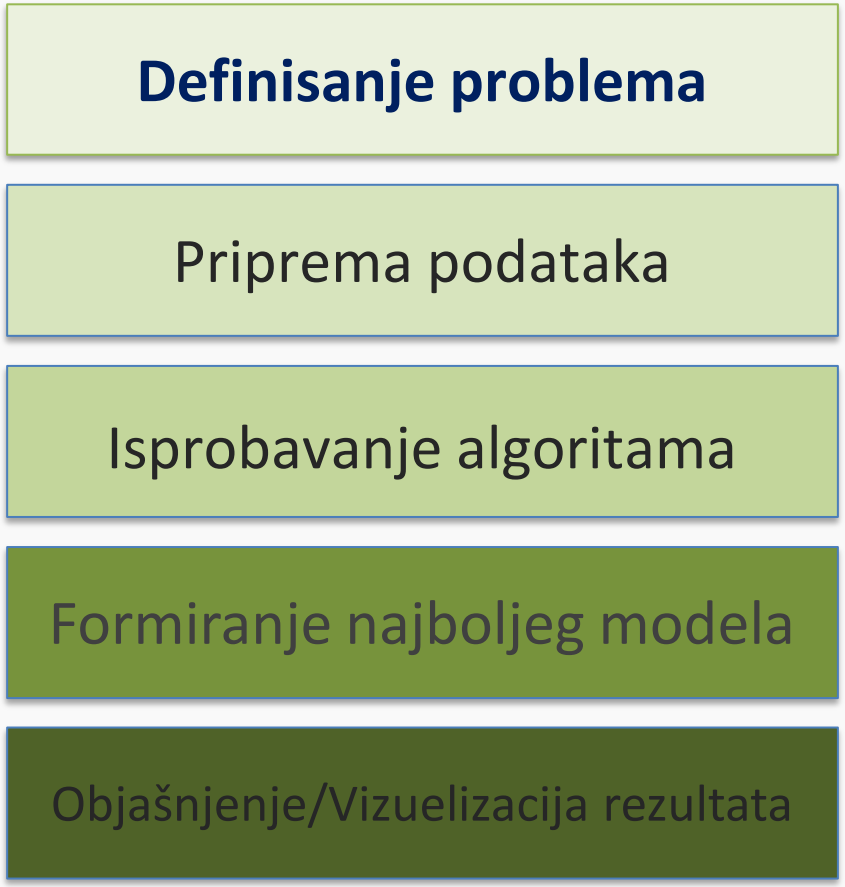


http://karpathy.github.io/2016/05/31/rl/

Koraci nadgledanog mašinskog učenja



Koraci nadgledanog mašinskog učenja



1. **Definisanje problema**

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

Formalni opis problema

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

T - zadatak

E - iskustvo koje se stiče treningom

P – mera performanse

Primeri

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.

Primer

Neformalni opis problema*: Potreban nam je program koji će moći da odredi koji tekstualni post na fejsu će biti deljen, na osnovu njegovog sadržaja.*

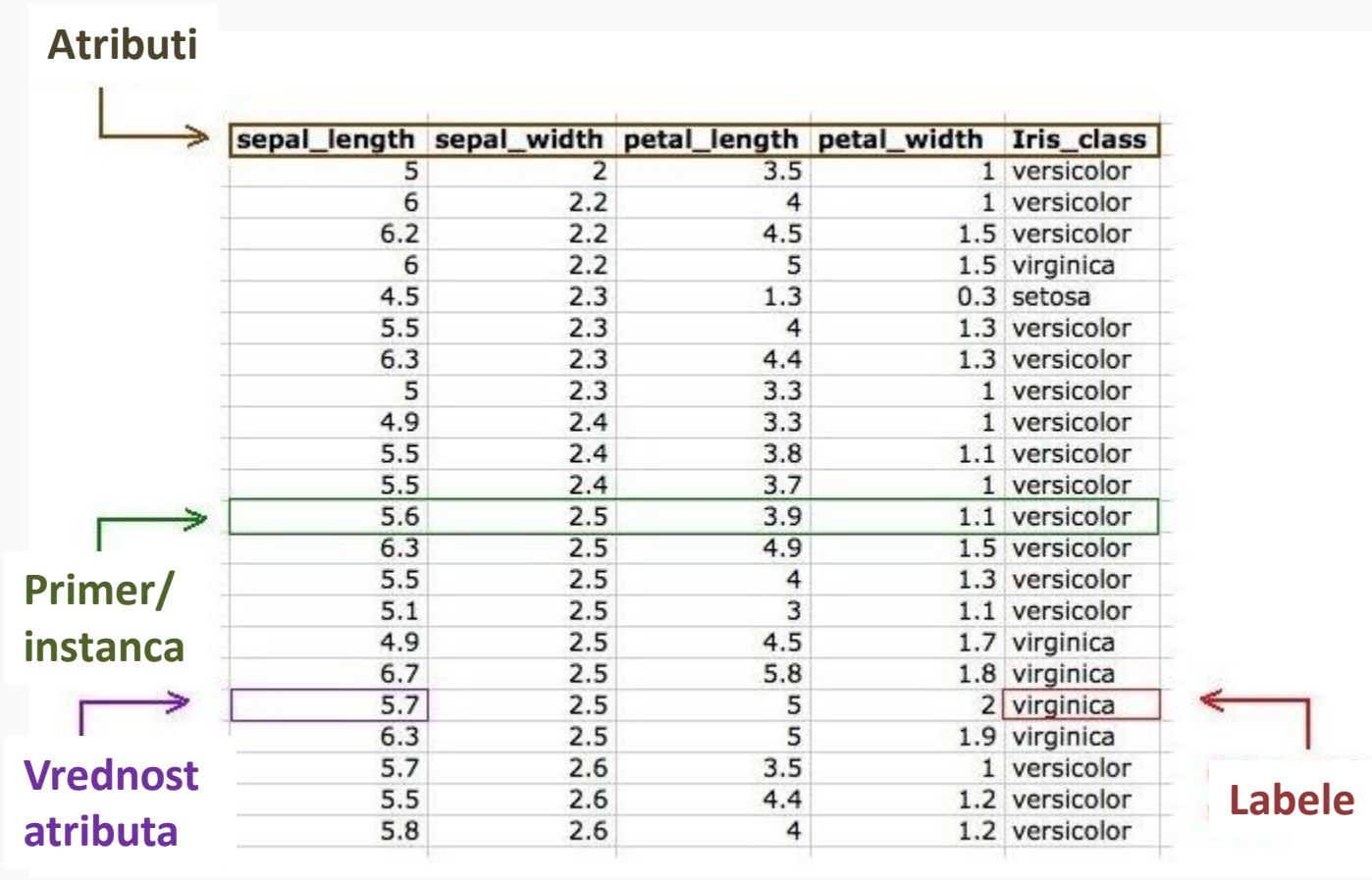
T – Klasifikacija posta koji nije objavljen (deli će se, neće se deliti)

E - Korpus postova među kojima ima onih koji nisu dalje deljeni i onih koji jesu.

P – Tačnost klasifikacije (broj ispravno klasifikovanih u ukupnom broju izvršenih klasifikacija)

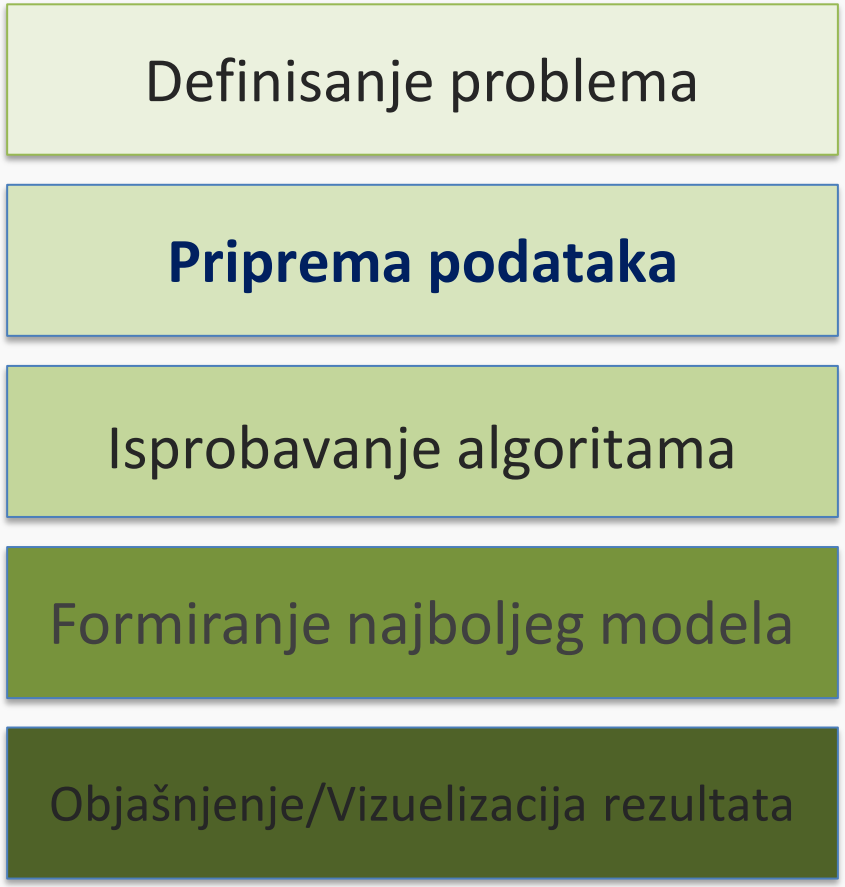
Koraci nadgledanog mašinskog učenja

- Definisanje problema
- **Priprema podataka**
- Isprobavanje algoritama
- Formiranje najboljeg modela
- Objašnjenje/Vizuelizacija rezultata



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

Koraci nadgledanog mašinskog učenja



1. **Priprema podataka**

Analiza
Odabir
Pretprocesiranje
Transformisanje

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

Primer

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}

Pogled na podatke

```
# 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

Pregled po atributima

```
# look at the features
heart_df.info();
```

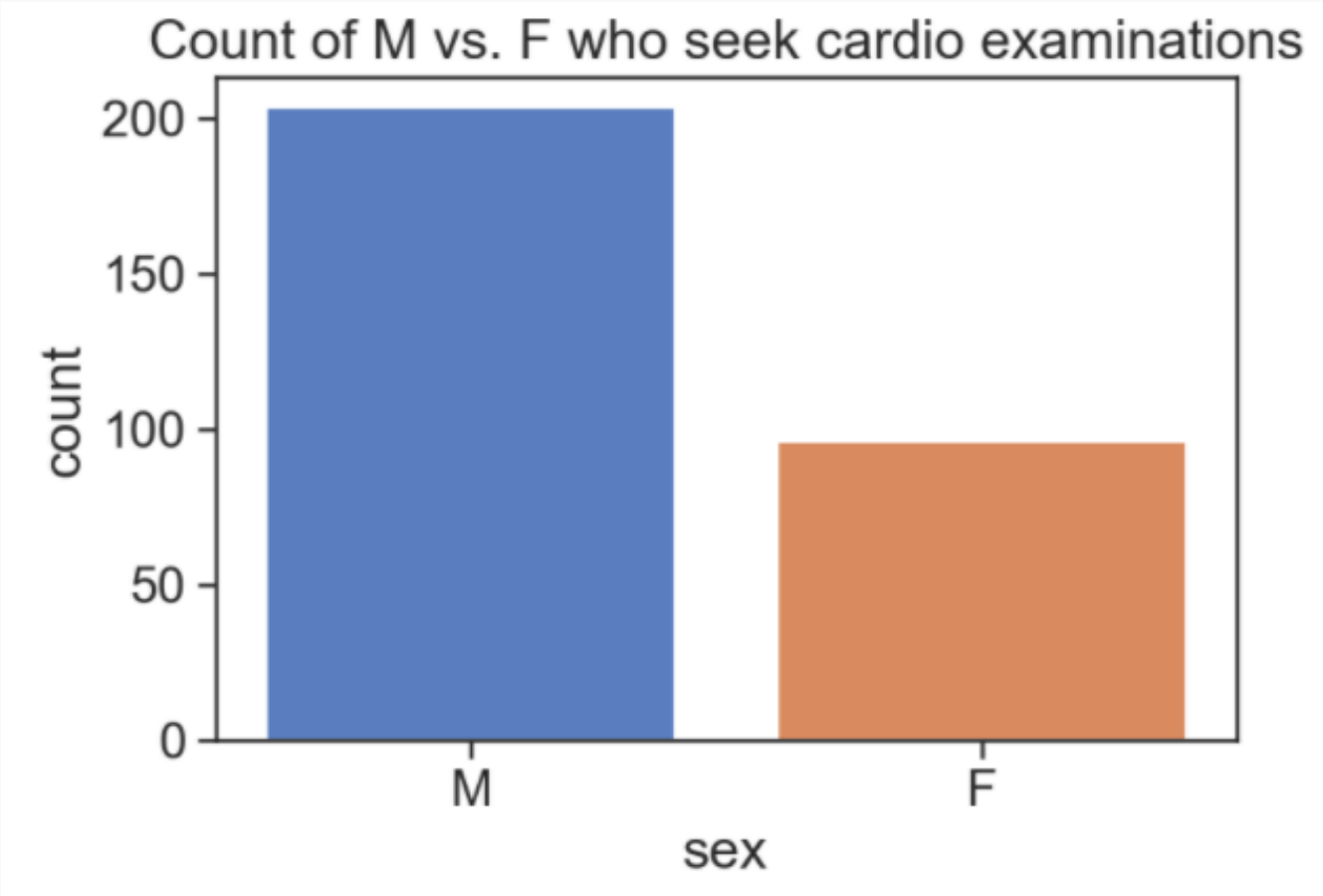
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 14 columns):
age         299 non-null float64
sex         299 non-null float64
cp          299 non-null float64
restbp      299 non-null float64
chol        299 non-null float64
fbs         299 non-null float64
restecg     299 non-null float64
thalach     299 non-null float64
exang       299 non-null float64
oldpeak     299 non-null float64
slope       299 non-null float64
ca          299 non-null float64
thal        299 non-null float64
hd          299 non-null float64
dtypes: float64(14)
```

Analiza podataka

1. At what ages do people seek cardiological exams?

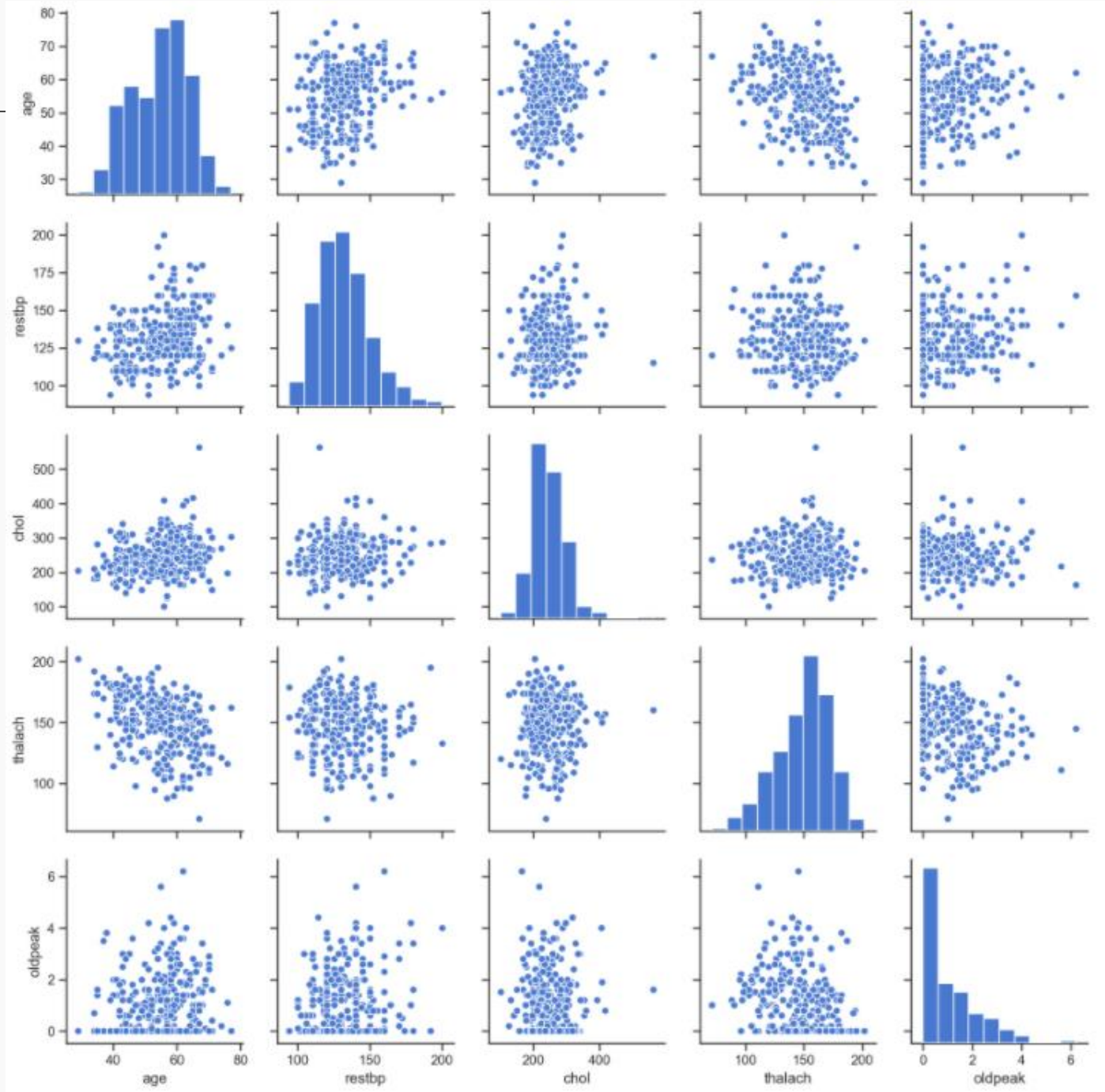
Analiza podataka

2. Do men seek help more than women?

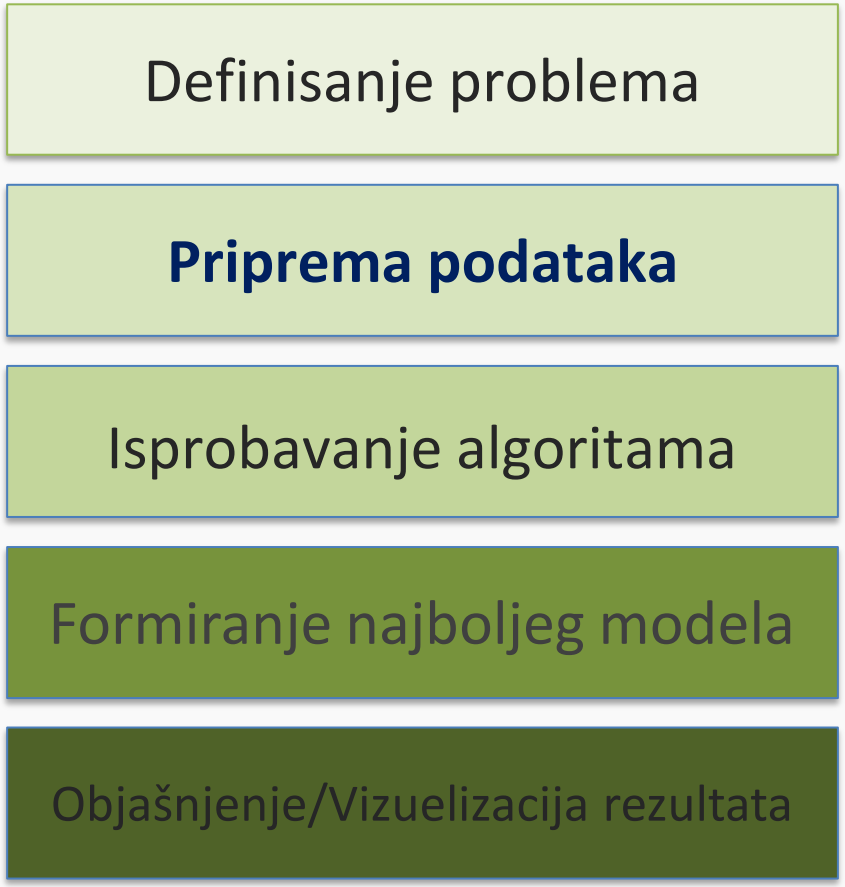


Analiza podataka

3. Examine the variables. How do they relate to one another?



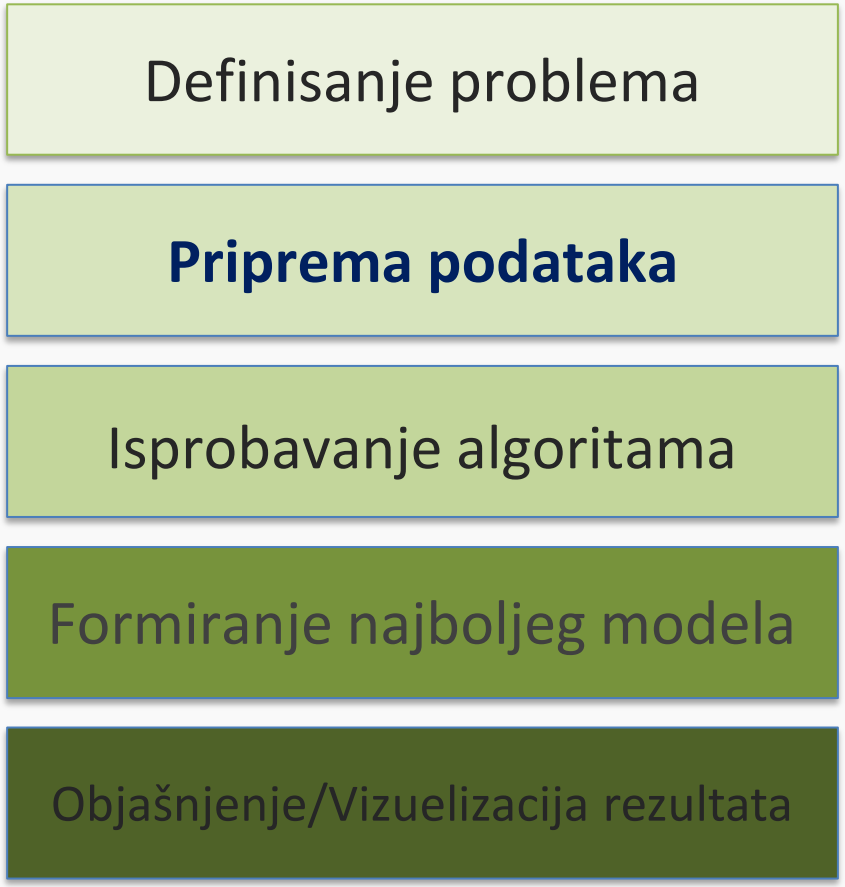
Koraci nadgledanog mašinskog učenja



1. **Priprema podataka**

Analiza
Odabir
Pretprocesiranje
Transformisanje

Koraci nadgledanog mašinskog učenja

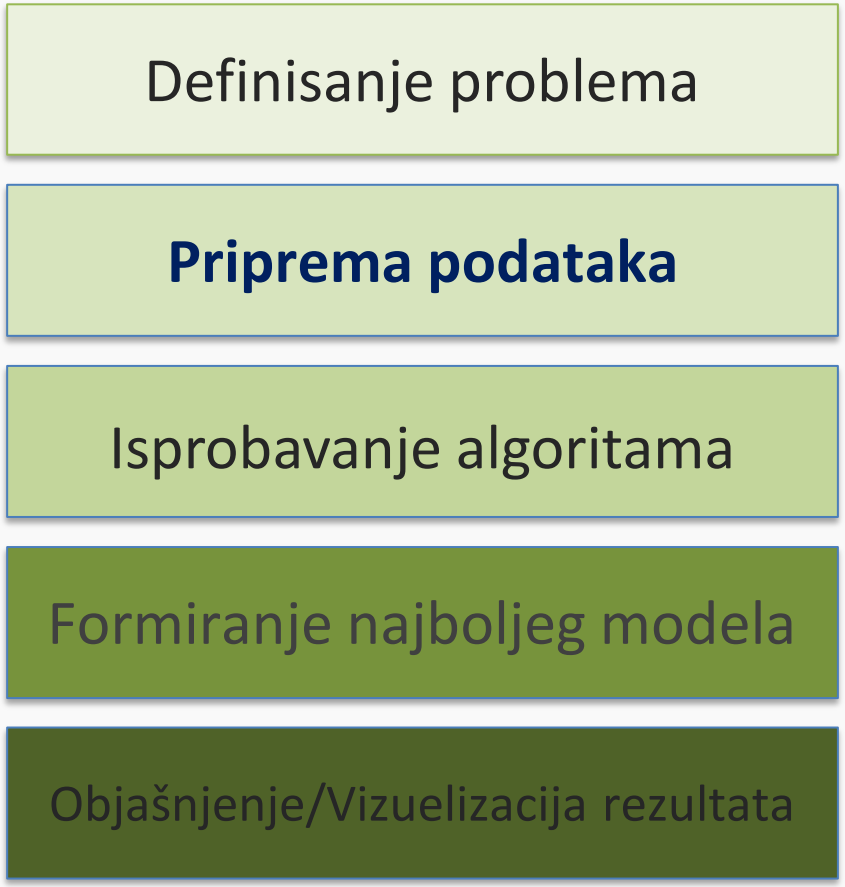


1. **Priprema podataka**

Analiza
Odabir
Pretprocesiranje
Transformisanje

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

Koraci nadgledanog mašinskog učenja



1. **Priprema podataka**

Analiza
Odabir
Pretprocesiranje
Transformisanje (Feature engineering)

- Skaliranje (scaling)
- Razlaganje atributa (attribute decomposition)
- Spajanje atributa (attribute aggregations)

Koraci nadgledanog mašinskog učenja

Definisanje problema

Priprema podataka

Isprobavanje algoritama

Formiranje najboljeg modela

Objašnjenje/Vizuelizacija rezultata

Šta je „algoritam“ u mašinskom učenju?

Postupak koji se izvodi nad podacima da bi se stvorio „model mašinskog učenja“.

Algoritmi mašinskog učenja vrše prepoznavanje uzorka u podacima (pattern recognition).

Algoritme za klasifikaciju (stabla odlučivanja). Algoritmi za regresiju, (Linearna regresija). Algoritmi za klasterizaciju (k-means).

Svojstva algoritama

Algoritmi mašinskog učenja mogu se opisati pomoću matematike i pseudokoda.

Efikasnost algoritama mašinskog učenja može se analizirati i opisati.

Algoritmi mašinskog učenja mogu se implementirati u bilo kom savremenom programskom jeziku.

Šta je „model“ u mašinskom učenju?

Izlaz iz algoritma mašinskog učenja pokrenutog nad podacima.

Rezultat učenja algoritmom mašinskog učenja.

Ono što ostaje sačuvano nakon pokretanja algoritma mašinskog učenja nad trening podacima.

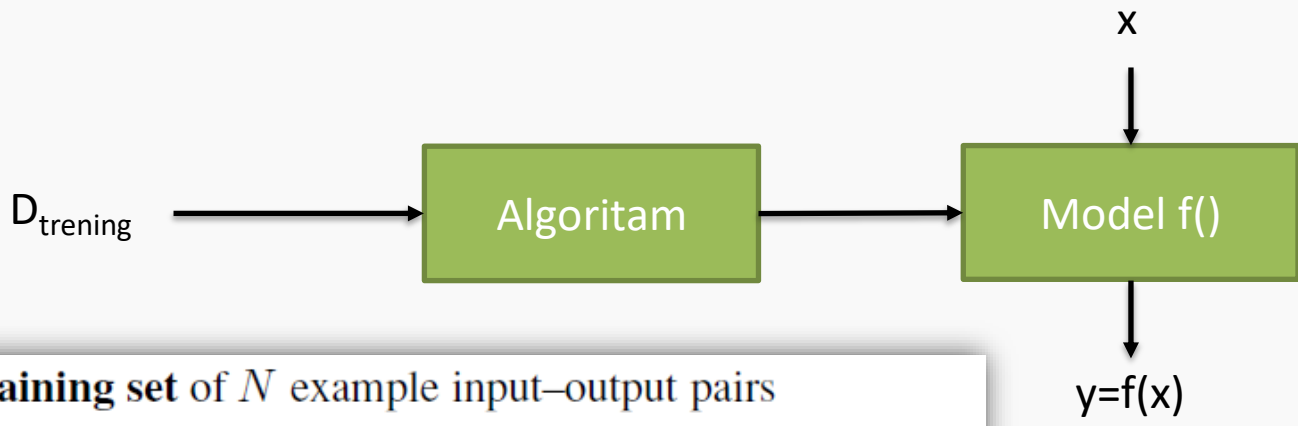
Predstavlja pravila, brojeve i sve druge algoritamski-specifične strukture podataka koje su potrebne za predikciju.

Učenje modela

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

f – model, hipoteza, prediktor

Algoritam uči funkciju f iz trening podataka.

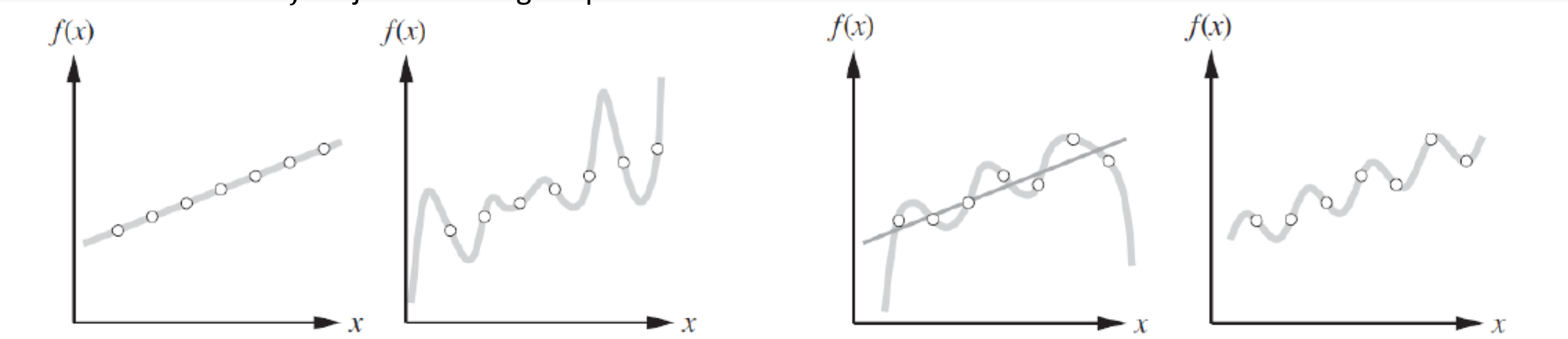


Given a **training set** of N example input–output pairs

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

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:



Koraci nadgledanog mašinskog učenja

Definisanje problema

Priprema podataka

Isprobavanje algoritama

Formiranje najboljeg modela

Objašnjenje/Vizuelizacija rezultata

Odabir najboljeg algoritma

Učenje predstavlja pretragu prostora mogućih modela, u cilju pronalaženja onog modela koji će imati dobre performanse i nad primerima koji nisu bili deo skupa za učenje.

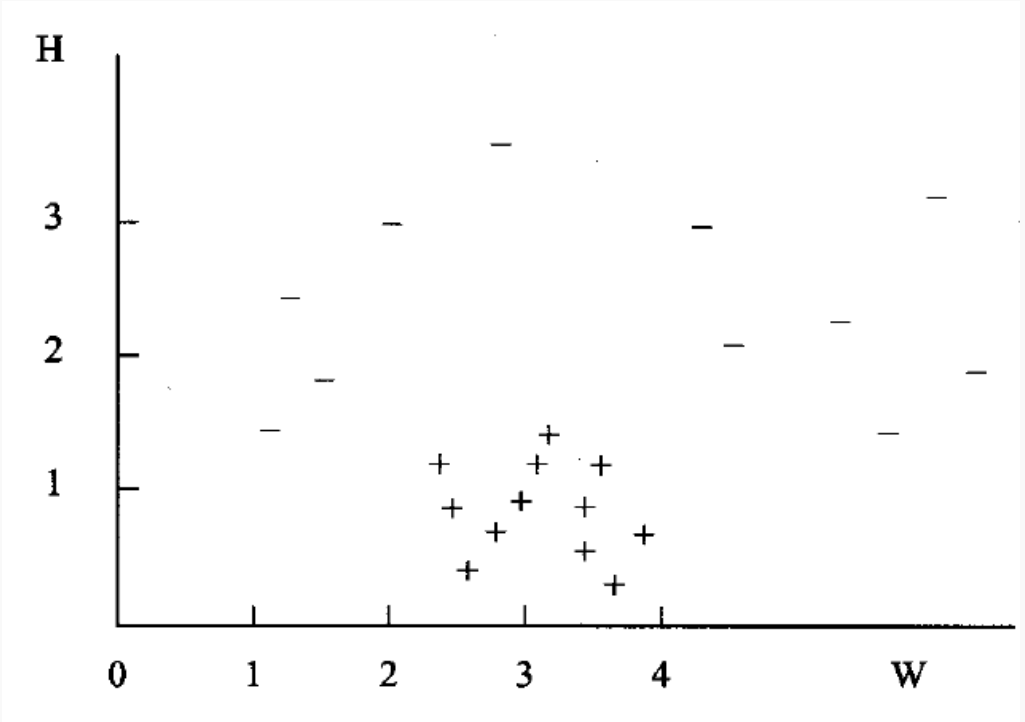
Kvalitet modela ocenjujemo korišćenjem skupa primera koji nisu bili deo skupa za učenje – **testni skup**.

Model dobro **generalizuje** ako korektno predviđa vrednosti *Y* za vrednosti *X* iz testnog skupa.

Primer različitih modela

Sakupili smo podatke o visini i širini mnogo različitih pečuraka i ekspert je za svaku od njih odredio da li je otrovna ili ne.

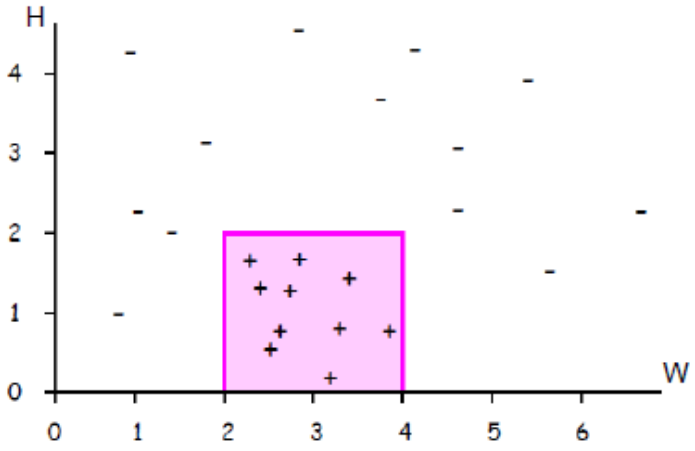
Skup za učenje je prikazan na slici tako što su obeležene jestive (‘+’) i otrovne pečurke (‘-’)



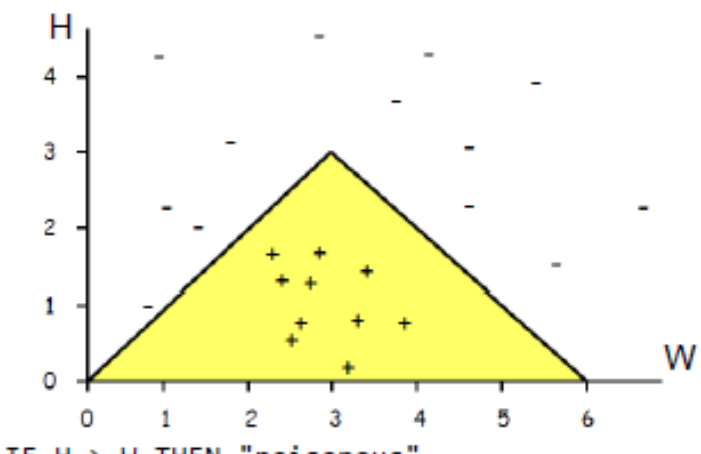
Primer različitih modela

Rezulat učenja je model - klasifikator koji je u stanju da klasifikuje nove pečurke

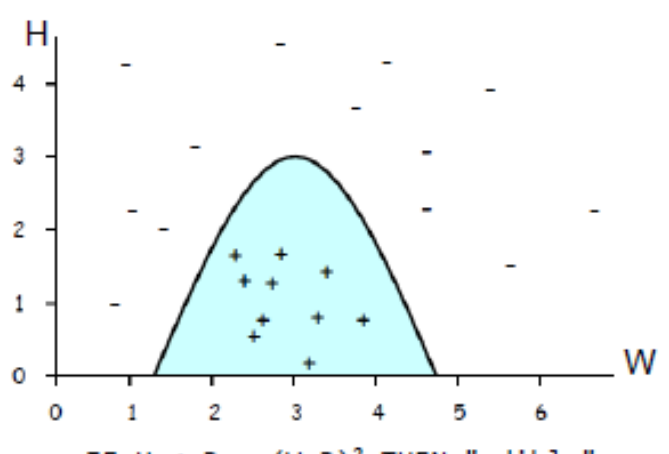
Klasifikator može biti dat i u formi if-then pravila:



IF W>2 and W<4 and H<2
THEN "edible" ELSE "poisonous"



IF H > W THEN "poisonous"
ELSE IF H > 6 – W THEN "poisonous" ELSE "edible"

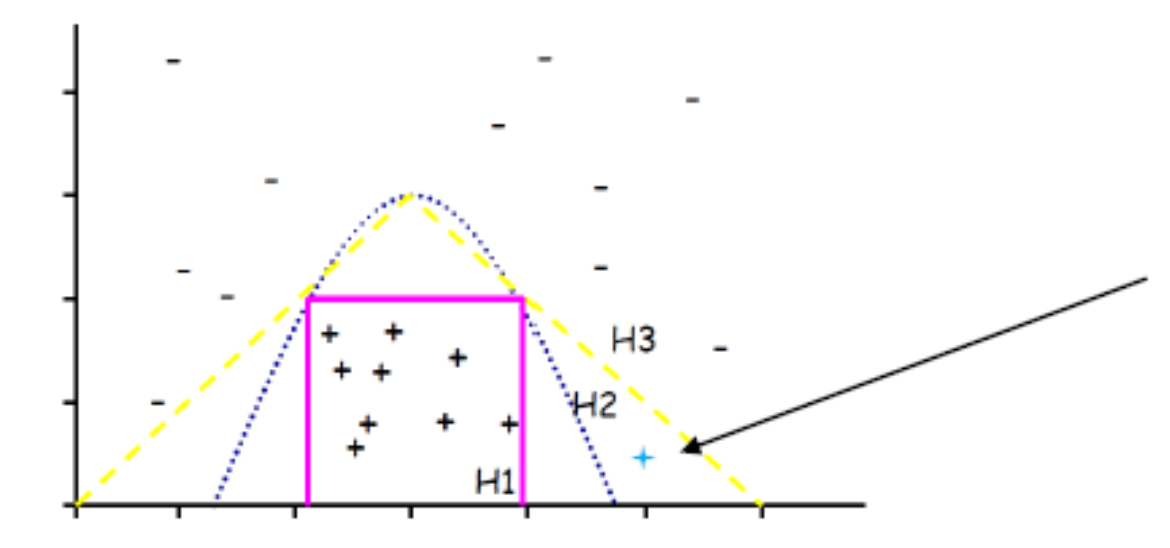


IF H < 3 – $(W-3)^2$ THEN "edible"
ELSE "poisonous"

Primer različitih modela

Svi modeli su u skladu sa skupom za učenje.

Šta je sa njihovim predikcijama na testnom skupu?



Kako klasifikujemo novi primer?

Koraci nadgledanog mašinskog učenja

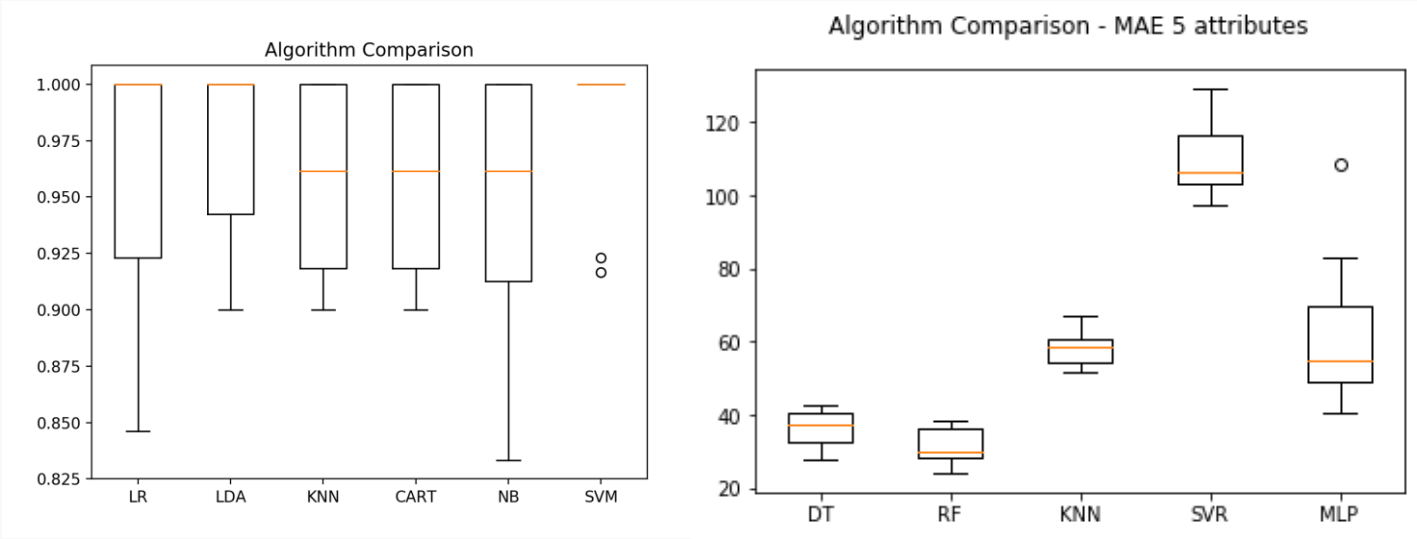
Definisanje problema

Priprema podataka

Isprobavanje algoritama

Formiranje najboljeg modela

Objašnjenje/Vizuelizacija rezultata

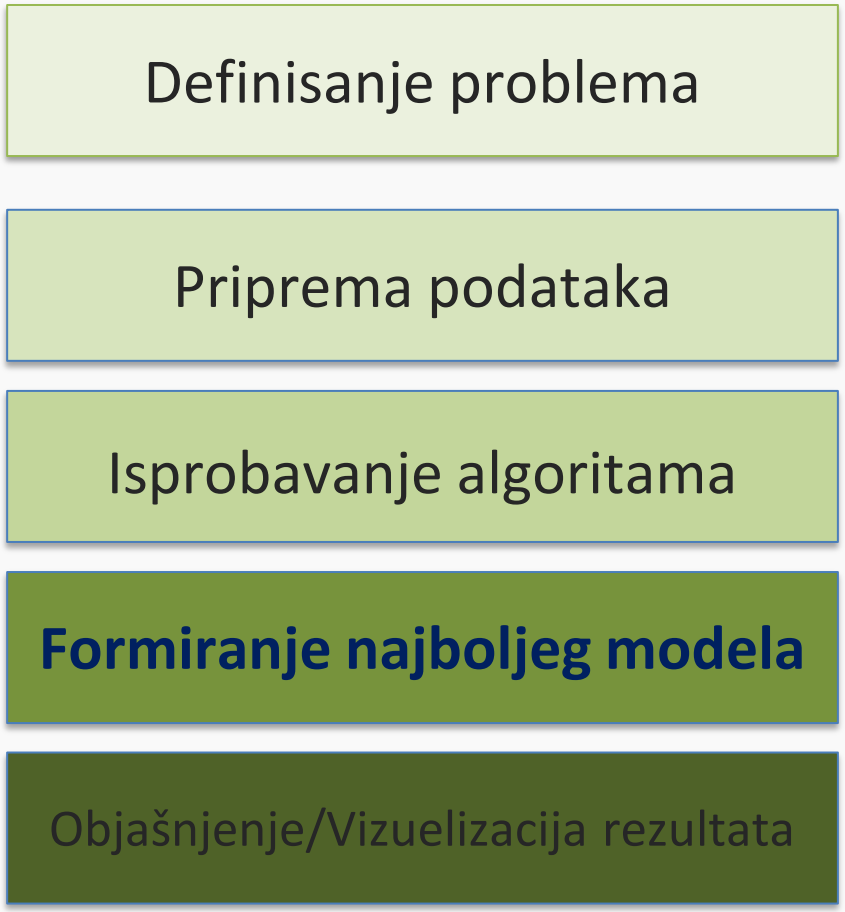


On a new problem, you need to quickly determine which type or class of algorithms is good at picking out the structure in your problem and which are not.

Koraci nadgledanog mašinskog učenja

- Podešavanje parametara algoritma (parameter tuning)
- Izgradnja ansambl modela
- Ponovno transformisanje podataka

Koraci nadgledanog mašinskog učenja



- Podešavanje parametara algoritma (parameter tuning)
- Izgradnja ansambl modela
- Ponovno transformisanje podataka

Koraci nadgledanog mašinskog učenja

Definisanje problema

Priprema podataka

Isprobavanje algoritama

Formiranje najboljeg modela

Objašnjenje/Vizuelizacija rezultata