Machine Learning

Ing. Silvia Komara, PhD.
Ekonomická univerzita v Bratislave

Obsah

- Úvod do Machine Learning
- Učenie s učiteľom (Supervised learning)
  - Klasifikácia
  - Regresia
- Učenie bez učiteľa (Unsupervised learning)
  - Redukcia dimenzionality
  - Detekcia anomálií
  - Zhlukovanie

Machine Learning (ML) – strojové učenie

Definícia

- veda o schopnosti počítačov učiť sa bez toho aby boli explicitne naprogramované (Arthur Samuel, 1959)
- Stojí na pilieroch štatistiky, informatiky a matematickej optimalizácie
- Je súčasťou Data Science (Dátovej vedy)
- Jeden zo spôsobov, akým sa snažíme dosahovať AI (Artificial Intelligence – umelá inteligencia)
- Hlavný zámer: aby sa počítač dokázal učiť samostatne, bez ľudskej asistencie a aby to čo sa naučil, vedel aj prakticky aplikovať

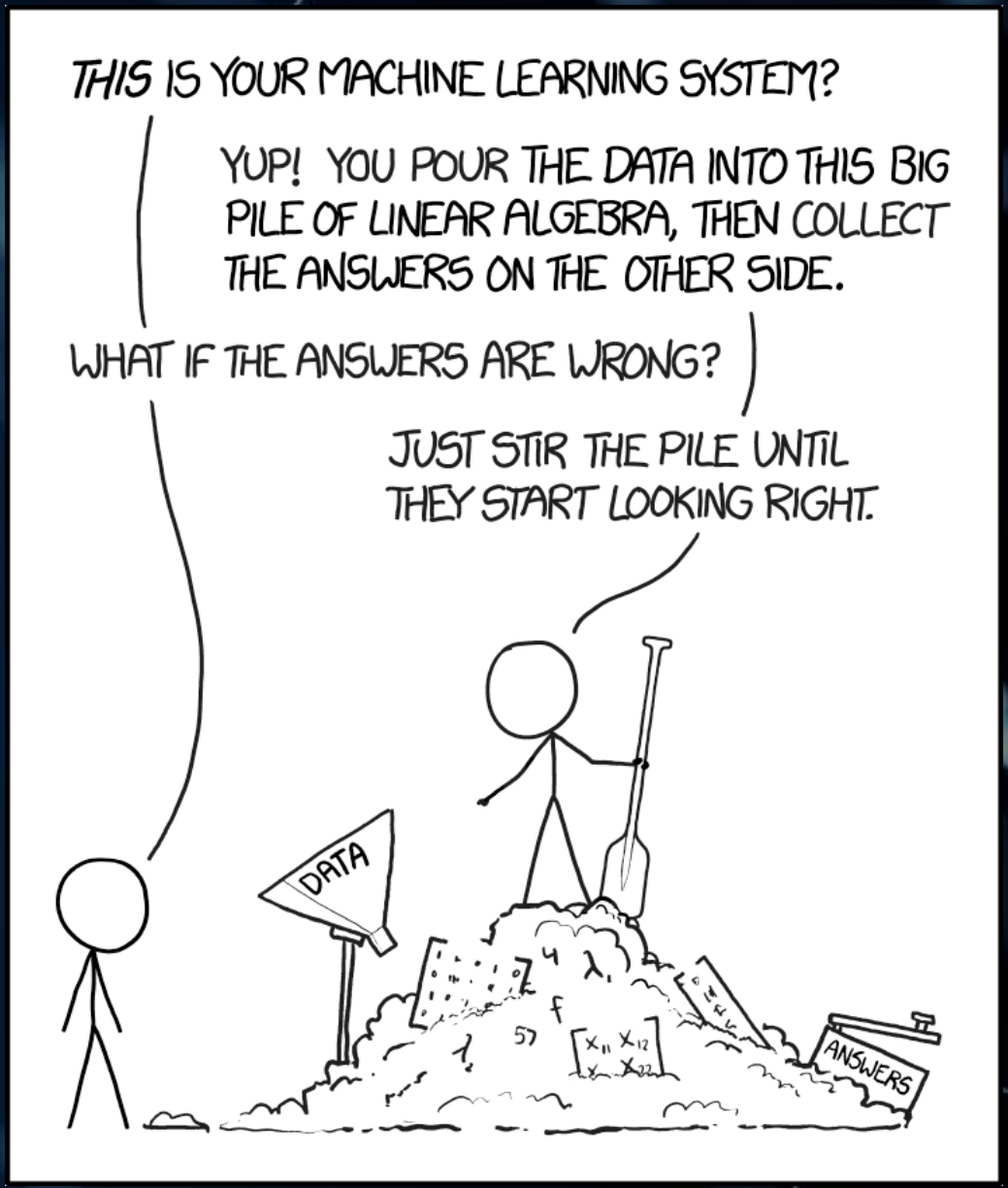
Príklady využitia

- Spam filter
- Automatické označovanie urážlivých komentov v diskusných fórach
- Segmentácia zákazníkov
- Odporúčanie filmov (Netflix) , pesničiek (Spotify), jedla, produktov a podobne.
- Označovanie priateľov na fotkách (FB, mobil, …)
- Chatbot (automatický asistent)
- Translátory
- Detekcia podvodov na bankových účtoch
- Medicína – detekcia nádorov a podobne.

Rozdiel medzi ML a štatistikou

- Cieľ
- Terminológia
- Softvér
- Veľkosť dát
- Výpočtový výkon
- Interpretácia
- Black box (čierna skrinka)

Príklad Black Box



Zdroj: https://imgs.xkcd.com/comics/machine_learning.png

Typy učenia

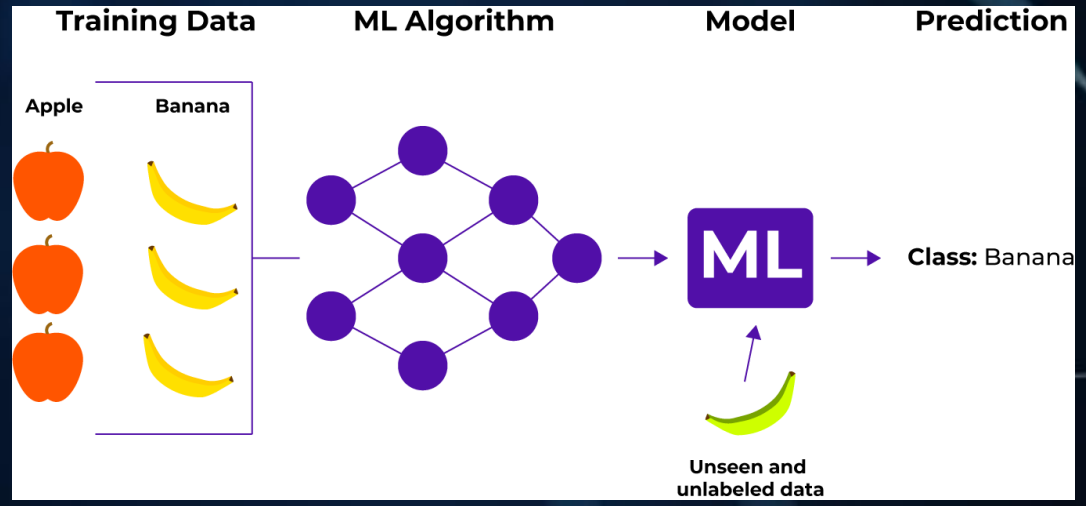
Na základe toho, ako konkrétne proces učenia prebieha, môžeme algoritmy strojového učenia rozdeliť do nasledujúcich troch skupín:

1. Učenie s učiteľom (Supervised machine learning)
2. Učenie bez učiteľa (Unsupervised machine learning)
3. Učenie formou odmeňovania (Reinforcement learning)

Učenie s učiteľom

- Najčastejšie používaný algoritmus ML
- Predpokladom pre jeho použitia je existencia dostatočne veľkého trénovacieho datasetu, ktorý obsahuje správne označené dvojice vstup – výstup (input – output/feature-label).

- Typy učenia s učiteľom:
  Klasifikácia - máme konečný počet kategórií (tried) a pri príchode nového vstupu chceme zaradiť, do ktorej kategórie patrí.
  Regresia - používa sa na predpovedanie, resp. odhad hodnoty nejakej spojitej premennej.

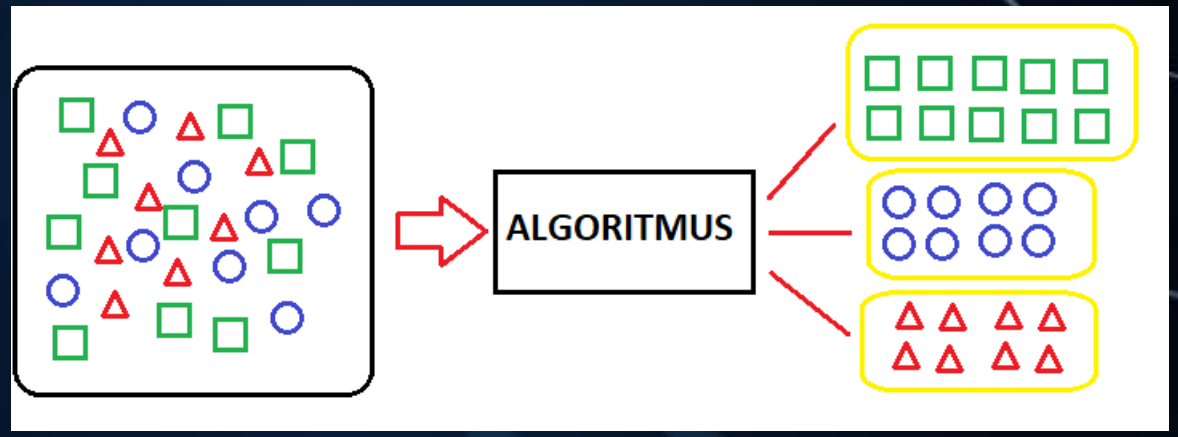


Zdroj: https://neurospace.io/blog/2020/08/what-is-supervised-learning/

Učenie bez učiteľa

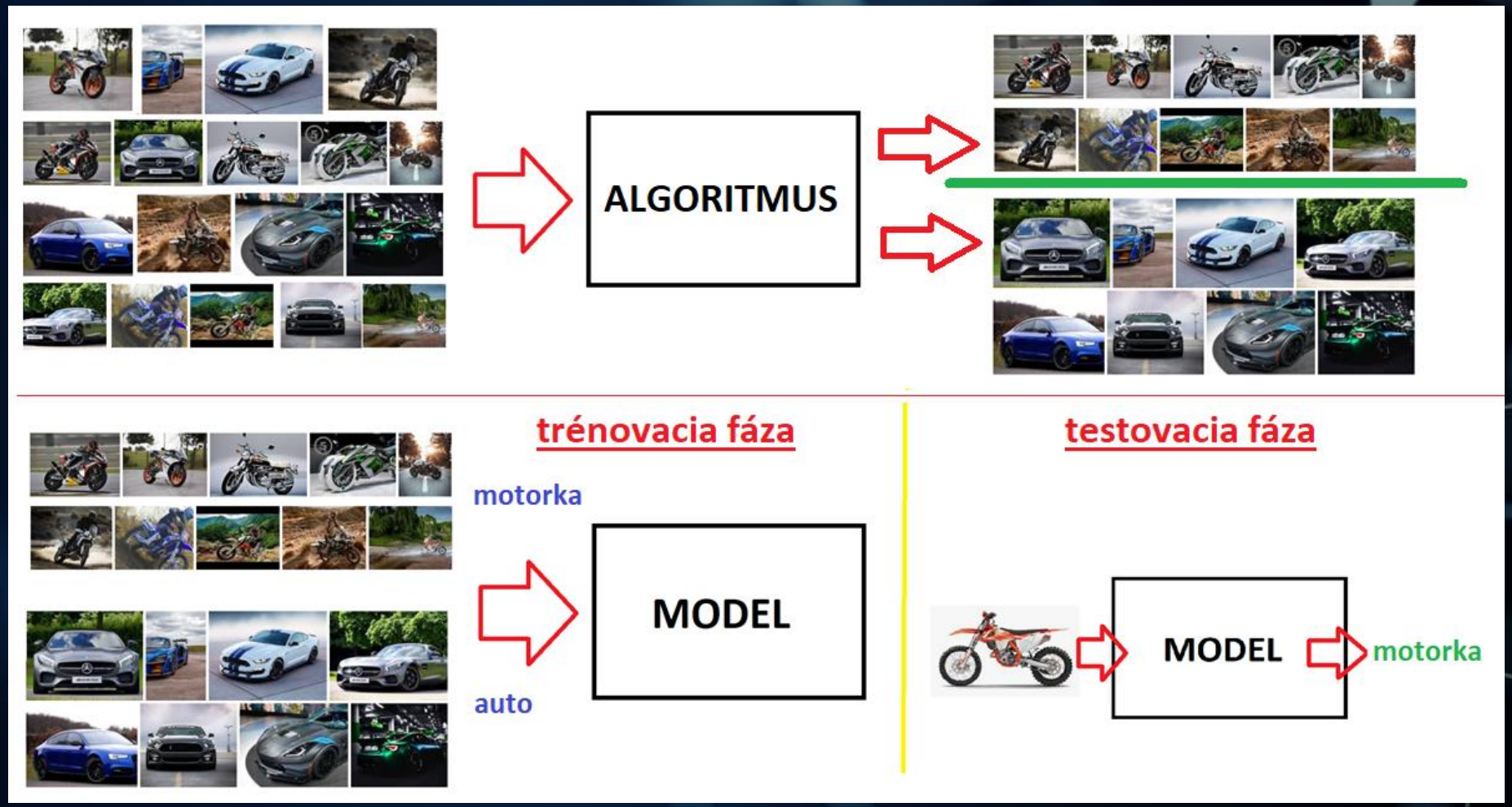
- Dáta nie sú označené (unlabeled)
- Algoritmy sa snažia objaviť, modelovať a popísať vzory v dátach, s cieľom dozvedieť sa niečo viac
- Popísanie skrytých štruktúr v dátach
- Najpoužívanejším algoritmom je ***zhlukovanie*** – spája dáta do skupín, ktoré majú niečo spoločné

Príklad: segmentácia zákazníkov do skupín s podobnými preferenciami.

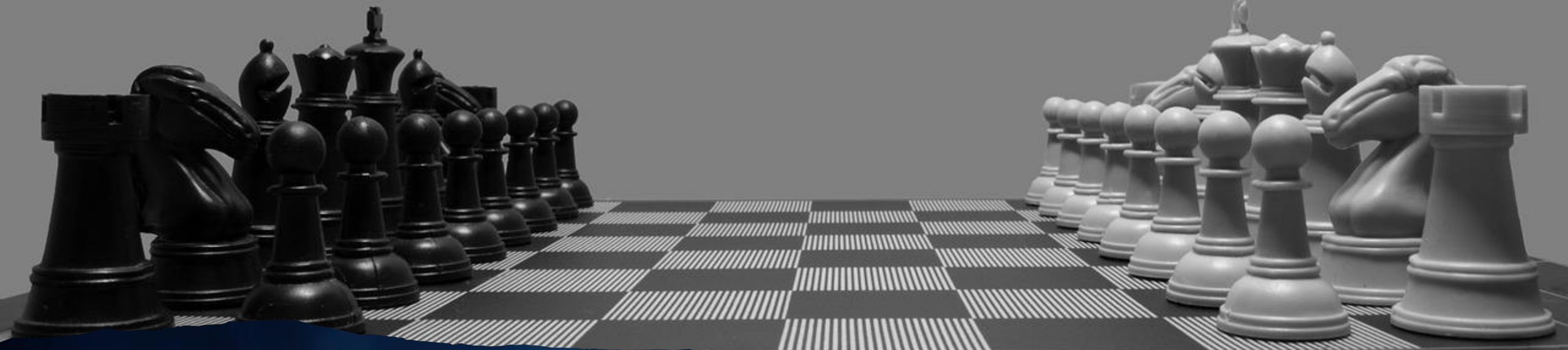


Zdroj: https://umelainteligencia.sk/algoritmy-strojoveho-ucenia-ii-ucenie-bez-ucitela/

Rozdiel medzi učením bez učiteľa (hore) a s učiteľom (dole)



Zdroj: https://umelainteligencia.sk/algoritmy-strojoveho-ucenia-ii-ucenie-bez-ucitela/



Učenie formou odmeňovania (Reinforcement learning)

- V tomto prípade už na trénovanie modelu nepoužijeme žiadne označené, či neoznačené trénovacie príklady. Učenie tu prebieha tak, že vytvoríme systém – agenta, ktorého nasadíme do prostredia a necháme ho nech sa učí prostredníctvom interakcie s prostredím.
- Jediné čo mu musíme určiť, sú pravidlá ako sa môže v danom prostredí správať a tzv. odmeňovaciu funkciu. Pomocou nej vie agent vyhodnotiť, či rozhodnutie, ktoré práve vykonal bolo preňho prospešné alebo nie. Následne metódou pokus-omyl podobne ako človek skúša jednotlivé možnosti a naučí sa ako sa má ideálne správať v jednotlivých situáciách. Modelovým príkladom pre učenie formou odmeňovania je hra šach (Deep blue vyhral nad Kasparovom, 1997).

Prečo Python??

- Najpopulárnejší programovací jazyk používaný na implementáciu ML a Data Science
- Jednoduchý a čitateľný
- Opensource, freeware
- Vopred pripravené knižnice pre ML a Deep Learning
- Masívna podpora komunity
- Najpopulárnejší programovací jazyk podľa TIOBE indexu, ktorý sa aktualizuje raz za mesiac. https://www.tiobe.com/tiobe-index/
- Druhý najžiadanejší programovací jazyk (Zdroj: https://www.devjobsscanner.com/)

Tvorca: **Guido van Rossum**
Dátum vzniku: **1991**

IDE pre Python

- Shell
- iPython
- Spyder (Anaconda)
- Jupyter Notebook
- PyCharm
- VS Code
- Google Colab
- Data Spell
- A ďalšie…

Knižnice Pythonu pre dátovú vedu a strojové učenie

- **NumPy** - práca s mnoho(viac)rozmernými dátovými poľami, lineárna algebra, pseudonáhodné generovanie čísel, lineárna regresia, logistická regresia, Naïve Bayes atď. https://numpy.org/
- **SciPy** – vedecká výpočtová knižnica. https://scipy.org/
- **Matplotlib** – najzákladnejší balík na vizualizáciu údajov v Pythone. https://matplotlib.org/
- **Seaborn** – knižnica vizualizácie dát založená na matplotlib. https://seaborn.pydata.org/
- **Pandas** – najpopulárnejšia knižnica na spracovanie a anlýzu dát. https://pandas.pydata.org/
- **Scikit-learn** – sa bežne používa pri riešení problémov strojového učenia a dátovej vedy. Obsahuje mnohé ML algoritmy. https://scikit-learn.org/stable/
- **TensorFlow** – knižnica, ktorá sa zameriava sa najmä na tréning a odvodenie hlbokých neurónových sietí . https://www.tensorflow.org/
- **Keras** – podmodul TensorFlow na vytváranie neurónových sietí na vysokej úrovni rozhrania. https://keras.io/
- **Pytorch** – používa sa na implementáciu hlbokého učenia a neurónových sietí na obrovských súboroch údajov. https://pytorch.org/

Tipy a triky

Chat GPT

Github Copilot

Supervised Learning
Postupnosť krokov

> *80 percent of a data scientist's valuable time is spent simply finding, cleansing, and organizing data, leaving only 20 percent to actually perform analysis...*
>
> IBM Data Analytics

1. Dáta
2. Exploratívna analýza
3. Čistenie a príprava dát
4. Definovanie cieľovej premennej (target/label) a nezávislých premenných (features)
5. Rozdelenie premenných na trénovaciu a testovaciu časť
6. Aplikácia algoritmu (.fit a .predict)
7. Vyhodnotenie úspešnosti modelu
8. Optimalizácia hyperparametrov

Supervised Learning
Klasifikácia

- **Klasifikácia** – snažíme sa predpovedať jednu z istého množstva tried

- Pri klasifikácii systém dostane vstupné vzory a informáciu o tom, do akej triedy patria. Cieľom je naučiť sa správne klasifikovať ľubovoľné ďalšie vstupné vzory – t.j., zatriediť ich do príslušnej triedy.

- **Typy klasifikácie**
  1. **Binárna (dichotomická) –** len dve možné cieľové triedy (áno/nie, 0/1 a podobne)
  2. **Viacnásobná –** viac ako dve cieľové triedy

Klasifikačné algoritmy

1. Rozhodovací strom (Decision Tree)
2. Náhodný les (Random Forest)
3. K-najbližší sused (KNN – *K*-Nearest Neighbors)
4. Lineárna diskriminačná analýza
5. Logistická regresia
6. Naïve Bayes
7. Metóda podporných vektorov (SVM - Support Vector Machines)
8. a ďalšie..

Klasifikácia
Hodnotenie kvality modelu

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- TP (True positive) a TN(True negative) – správne odhadnuté hodnoty
- FP (False positive) a FN (False negative) – nesprávne odhadnuté hodnoty

- Accuracy (úspešnosť) = $\frac{TP+TN}{TP+TN+FP+FN}$

(nie je vhodná pre nevyvážený dataset)

- Precision (presnosť) = $\frac{TP}{TP+FP}$

(úspešnosť predikcie jednej triedy)

- Recall (pokrytie) = $\frac{TP}{TP+FN}$

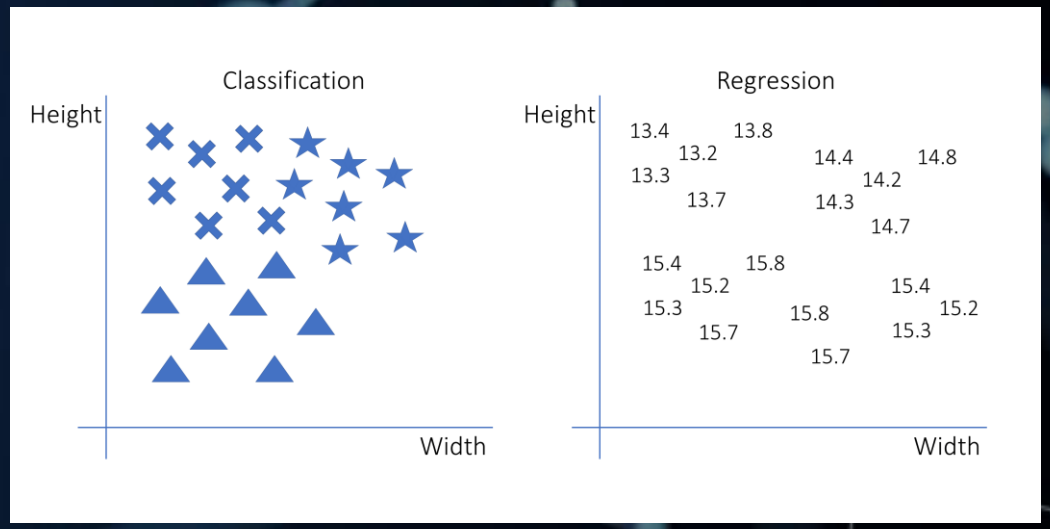
- F1-skóre = $2\,\frac{precision*recall}{precision+recall}$

Regresia

- Pre regresnú úlohu je cieľom predikovať číselnú premennú, napríklad cena bytu v závislosti od polohy bytu, rozlohy bytu, veku a podobne.

Algoritmy pre regresný problém:

1. Lineárna regresia
2. Hrebeňová regresia (Ridge regression)
3. LASSO regresia
4. Rozhodovací strom
5. Náhodný les
6. *K*-najbližší sused
7. Metóda podporných vektorov
8. Polynomická regresia
9. XGBoost



Zdroj: https://realpython.com/knn-python/

Regresia - vyhodnotenie kvality modelu

1. Miery presnosti prognóz:

- Rozptyl chýb MSE (Mean Square Error) $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Smerodajná odchýlka chýb prognóz RMSE (Root mean square error) $RMSE = \sqrt{MSE}$
- Miery presnosti sú založené na meraní chyby prognózy – rozdiel medzi skutočnou a odhadnutou hodnotou $e_i = y_i - \hat{y}_i$ (kde e_i nazývame rezíduum)
- Cieľom je minimalizovať chyby prognózy

2. Koeficient determinácie $r_{xy}^2 = (r_{xy})^2$ $r_{xy}^2 \in \langle 0,1 \rangle$

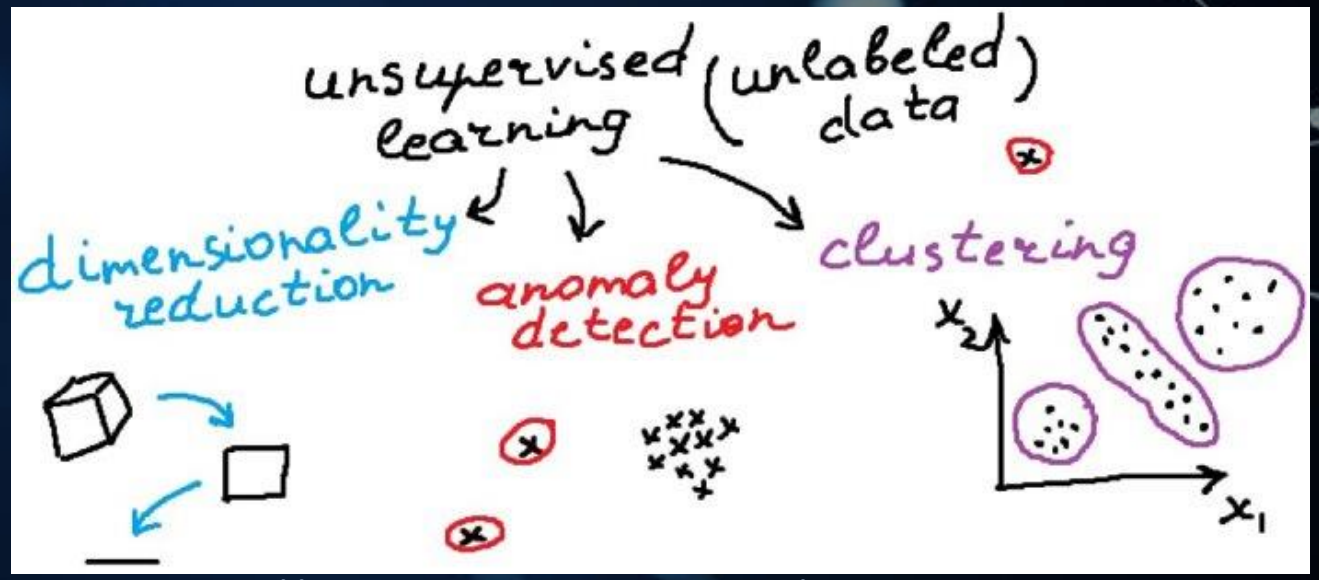
- vyjadruje kvalitu regresného modelu – interpretujeme ho v percentách
- udáva, koľko percent variability závisle premennej *Y* vieme vysvetliť pomocou daného odhadnutého regresného modelu. Zvyšná časť potom predstavuje percento nevysvetlenej variability, ktoré pripisujeme iným vplyvom alebo náhode.
- hodnoty blízke 0 predstavujú veľmi zlý model a hodnoty blízke k 1 značia dobrú kvalitu modelu.

Učenie bez učiteľa

(Unsupervised learning)

Učenie bez učiteľa (unsupervised learning)

- Používa sa pri dátach, ktoré neboli vopred označené. Chýba nám teda „správna odpoveď“.
- Môžeme mať totiž prípady, kedy túto správnu odpoveď nie je možné odpozorovať alebo ju ani sami nevieme jednoznačne určiť. Algoritmy učenia bez učiteľa sa potom v týchto neklasifikovaných dátach snažia objaviť, modelovať a popísať vzory, s cieľom dozvedieť sa o týchto dátach niečo viac.
- Algoritmus teda nepríde na správny výstup, ale preskúma dáta a popíše skryté štruktúry v týchto neoznačených dátach.



Zdroj: https://towardsdatascience.com/unsupervised-learning-algorithms-cheat-sheet-d391a39de44a

Yann LeCunn-ov koláč

„Ak by inteligenciou bol koláč, učenie bez učiteľa by bol koláč, učenie s učiteľom by bola poleva a učenie s odmeňovaním by bola čerešnička na torte.“

- Výrokom chcel vyzdvihnúť dôležitosť a obrovský potenciál algoritmov učenia bez učiteľa

- Yann LeCunn (narodený 8.6.1960 vo Francúzsku)
  - Vice-prezident a šéf AI v Meta
  - Profesor na New York University

1. Redukcia dimenzionality

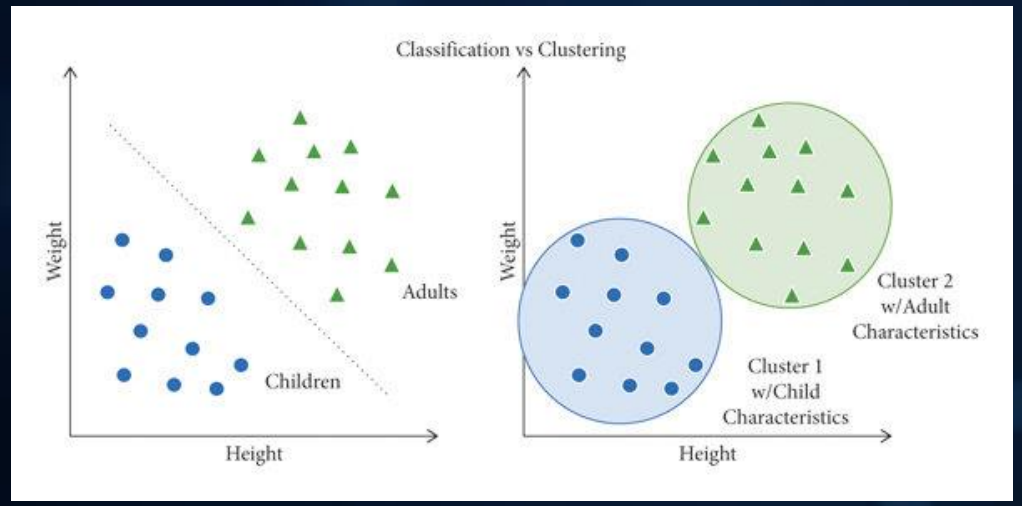
- Považuje sa za metódu predprípravy dát
- Cieľom je zníženie počtu vstupných premenných v súbore
- Algoritmy: **metóda hlavných komponentov** (PCA – Principal Component Analysis)**, nelineárna redukcia dimenzionlity** (Manifold Learning), **autoenkódery** (Autoencoders) – špeciálna architektúra umelých neurónových sietí

2. Detekcia anomálií (Anomaly/Outlier Detection)

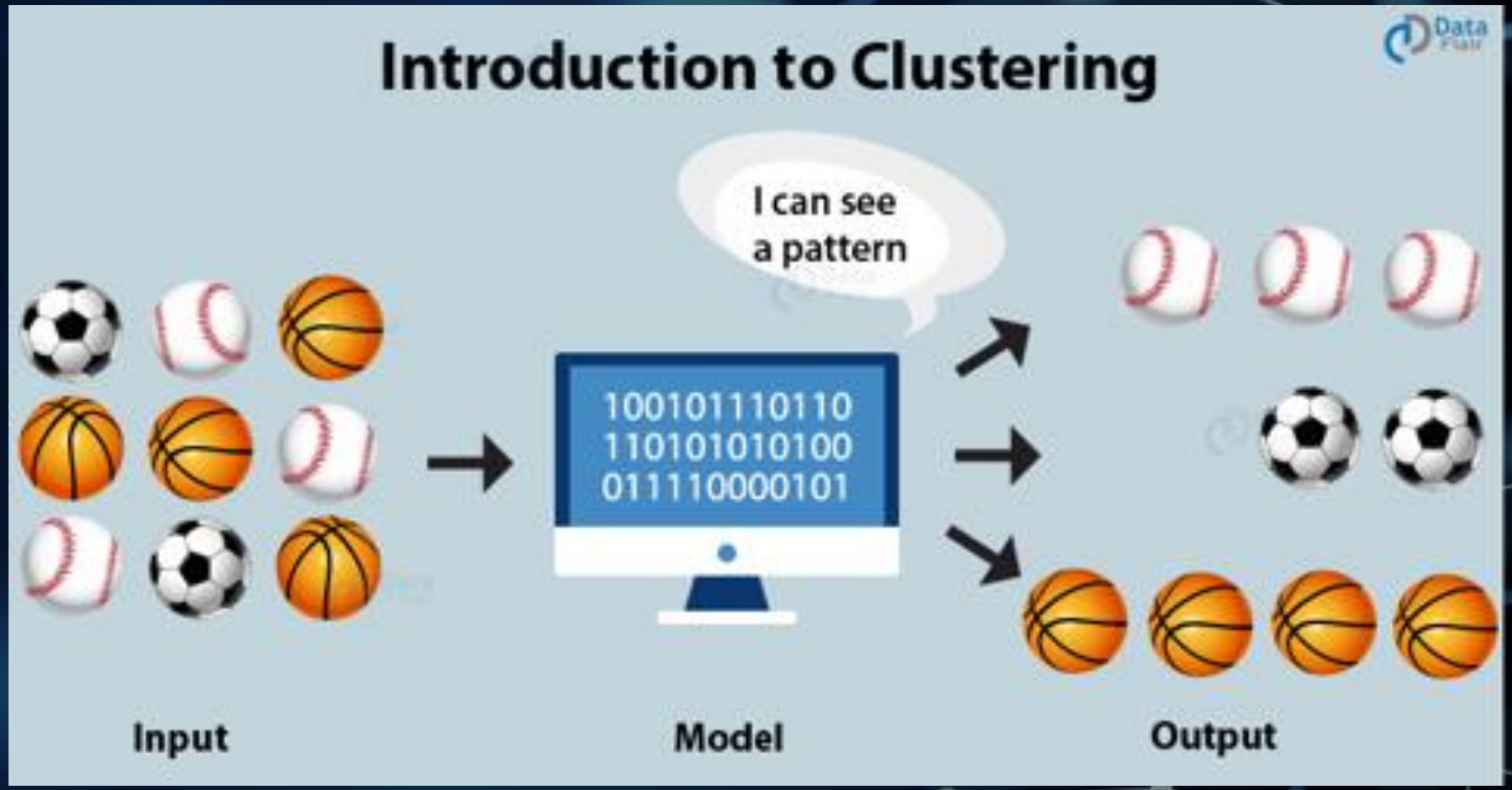
- Považuje sa za metódu čistenia a prípravy dát (Data Wrangling)
- Používa sa ale aj ako úloha na odhalenie podvovov (Fraud Detection)
- Algoritmy: **Tukey metóda medzikvartilového rozpätia, *z*-skóre, izolovaný les, SVM, lokálny odľahlý faktor** a iné.

3. Zhlukovanie (Clustering)

- Cieľom zhlukovania je rozklad súboru objektov na niekoľko relatívne rovnorodých podmnožín (zhlukov) tak, aby objekty patriace do rôznych zhlukov si boli čo najmenej podobné a objekty patriace do toho istého zhluku si boli podobné čo najviac t.j., algoritmus vytvára zhluky prvkov s podobnými vlastnosťami
- Ukážkovým príkladom je segmentácia zákazníkov do skupín s podobnými preferenciami.
- Zhlukovanie sa podobá klasifikácii, ale na rozdiel od nej zaraďuje prvky do tried bez mien



Zdroj: Hou, Shujun. (2021). Research on the Application of Data Mining Technology in the Analysis of College Students' Sports Psychology. Mobile Information Systems. 2021. 1-7. 10.1155/2021/6529174.



Zdroj: https://data-flair.training/blogs/clustering-in-machine-learning/

Algoritmy zhlukovania

1.Hierarchické zhlukovacie metódy – nevyžadujú apriori poznať počet zhlukov analýzy, čo je ich hlavná výhoda oproti nehierarchickým postupom. Nevýhodou je, že objekt, ktorý už bol zaradený do jedného zhluku na určitej úrovni zhlukovania, nemôže byť znovuzaradený do iného na ďalšej úrovni zhlukovania.

- **Metóda najbližšieho suseda**
- **Metóda najvzdialenejšieho suseda**
- **Metóda priemernej vzdialenosti**
- **Centroidná metóda**
- **Mediánová metóda**
- **Wardova metóda**

2.Nehierarchické zhlukovacie metódy – vychádzajú z požiadavky apriórnej znalosti o počte zhlukov na začiatku analýzy. Ide o iteračné postupy, ktoré sa obvykle začínajú zadaním počiatočného rozkladu súboru, ktorý sa postupne vylepšuje.

- **Metóda typických bodov**
- **Metóda *k*-priemerov**

Metóda *k*-priemerov (*K*- means)

•Je založená na centroidnom koncepte, kde **centroid** je ťažiskový bod/geometrický stred – priemer ukazovateľov v jednotlivých zhlukoch. Na začiatku sú centroidy umiestnené náhodne. Iteračným postupom sa poloha centroidov optimalizuje - posúvajú sa na základe vypočítaných nových stredných hodnôt centroidov. Na konci analýzy sa centroidy nachádzajú v strede konečných zhlukov.

•Počet počiatočných centroidov vypočítame na základe vnútrozhlukovej homogenity a **metódy lakťa** (**Elbow method**). **Mierou homogenity** (**inertia**) je **vnútrozhluková suma štvorcov odchýlok od priemeru zhluku** (**SSE** – sum of square error). Optimálny počet centroidov je tá hodnota, kde sa pokles SSE výrazne spomalí a od daného bodu klesá už len veľmi pomaly.

Postup:

1.Výber *k* počiatočných zhlukových centroidov

2.Zaradenie každého objektu do zhluku, ku ktorému je najbližšie (euklidovská vzdialenosť)

3.Premiestnenie centroidov na základe optimalizácie polohy a znovuzaradenie alebo premiestnenie každého objektu do jedného z *k* zhlukov

4.Ukončenie premiesťovania alebo znovuzaraďovania objektov, ak určené pravidlo dosahuje požadované kritérium (poloha centroidu je optimalizovaná).

K – Means Clustering

Zdroj: https://medium.com/@quantum_tunnel/getting-started-with-k-means-clustering-in-python-d4be91464cfb

Kontaktné údaje:

Ing. Silvia Komara, PhD., Katedra štatistiky, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35 Bratislava, e-mail: silvia.komara@euba.sk

Projekt:

VEGA c . 1/0431/22 – Implementácia inovatívnych prístupov modelovania rizík v procese ich riadenia v interných modeloch poisťovní v kontexte s požiadavkami direktívy Solvency II

Vypracované:

Január 2024