Adatbányászati módszerek

Kovács László, ME

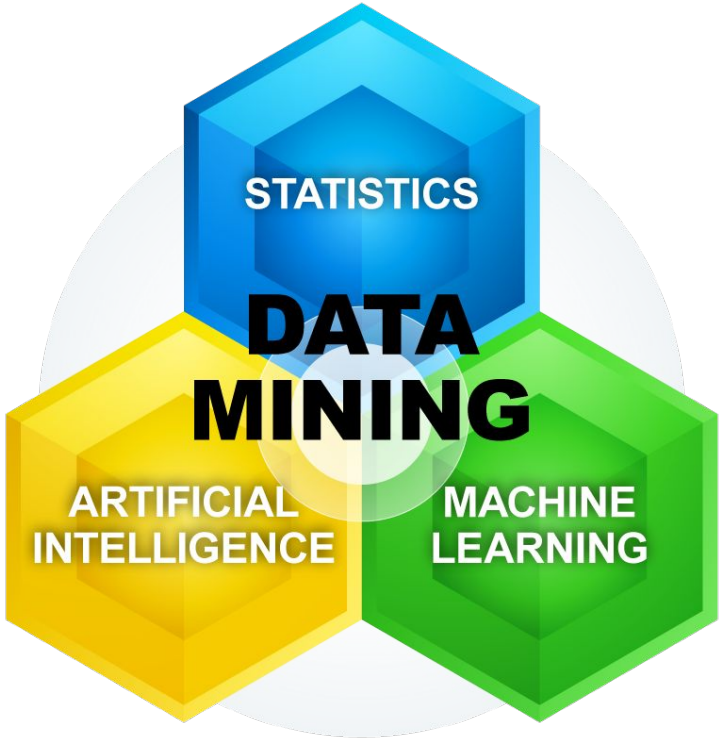
Adatbányászat

Szakértő rendszerek:

- szabály alapú döntések
- szakértő által megadott szabályok
- IF .. THEN .. alakú szabályok

Gépi tanulás (Machine Learning)

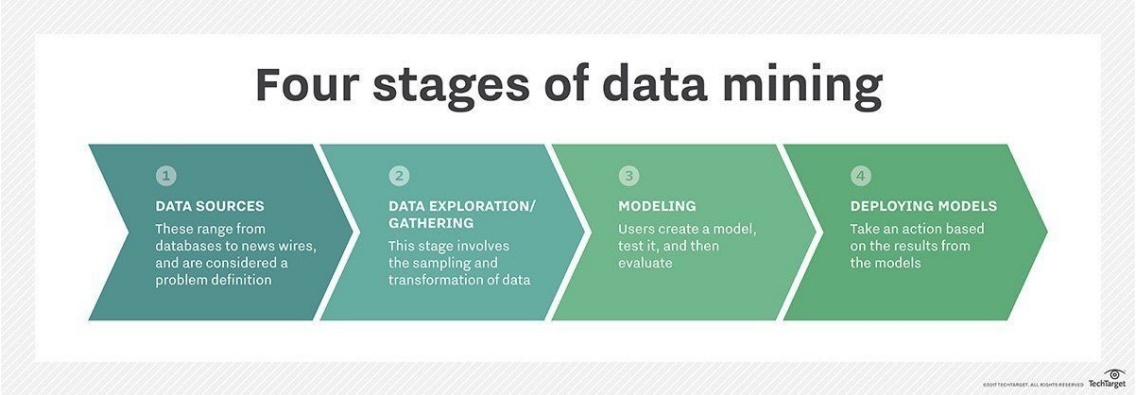
- implicit szabályok
- a szabályok a tanító adatokból állnak elő
- releváns szabályok keresése
- statisztika alapú döntések (legvalószínűbb viselkedés feltárása)



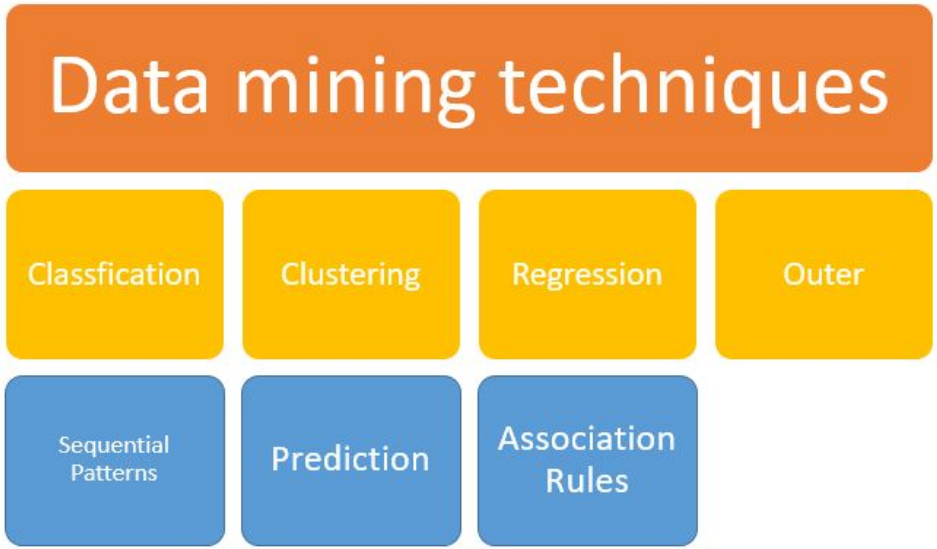
Adatbányászat

Feldolgozási lépések:

- cél és adatforrás meghatározása
- adatgyűjtés
- adat tisztítás
- adat redukció
- módszer kiválasztása
- paraméter megadása
- modellezés, tanítás
- predikció,
- értékelés
- alkalmazás
- finomítási ciklus



Data Mining

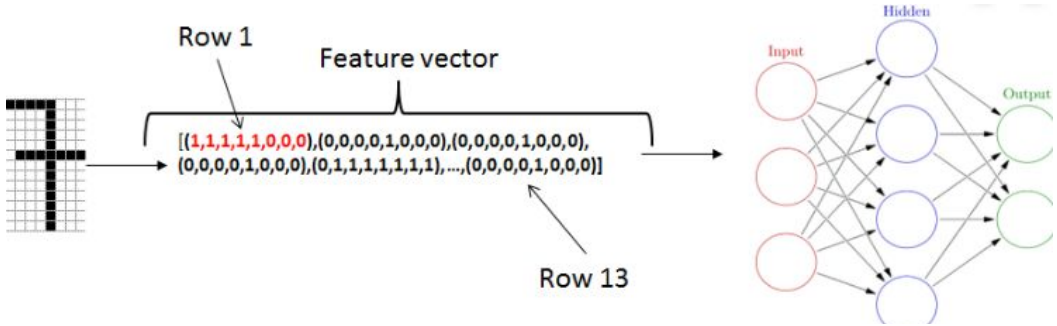

Data mining techniques
Classfication
Clustering
Regression
Outer
Sequential Patterns
Prediction
Association Rules

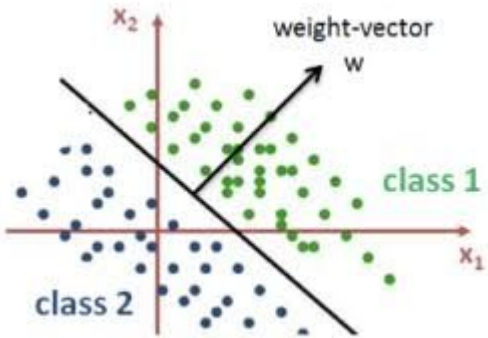

Database Marketing
Credit Risk Management
Training and Support
Fraud Detection
Healthcare Bioinformatics
Sentiment Analysis
Qualitative Data Mining
Spam Filtering

Adatreprezentáció

Tulajdonság vektor:

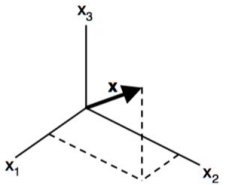
- vektortér modell
- tulajdonságok, dimenziók
- vektor műveletek
- univerzális műveletek
- ikiterjeszthető



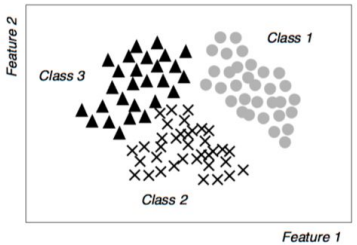


$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_d \end{bmatrix}$$

Feature vector



Feature space (3D)

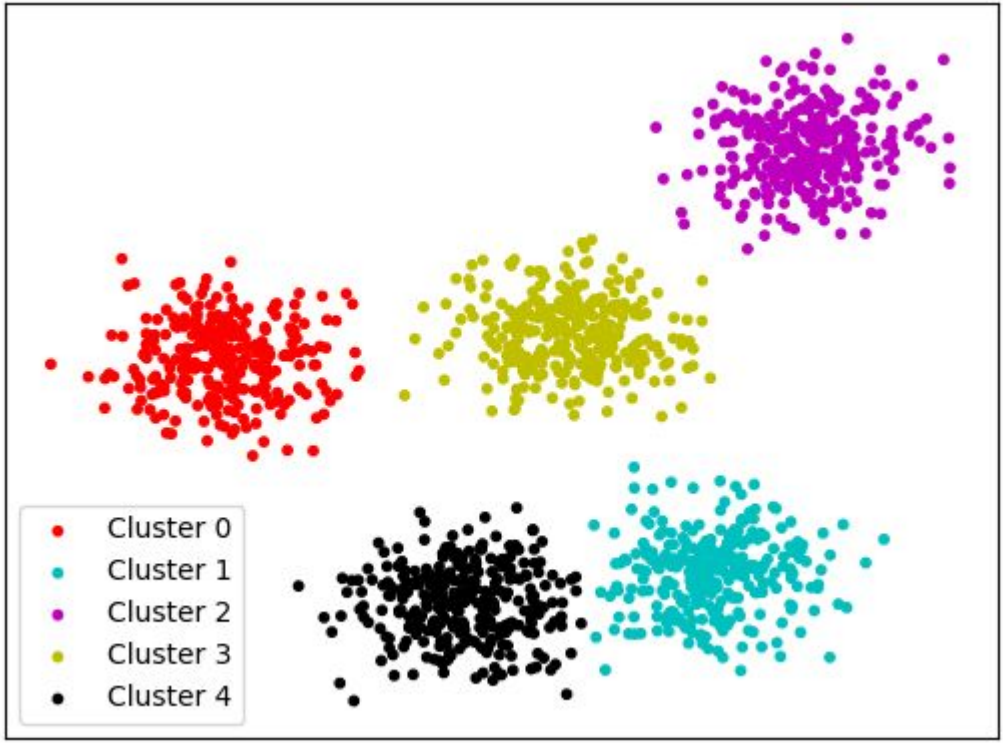


Scatter plot (2D)

Klaszterezés

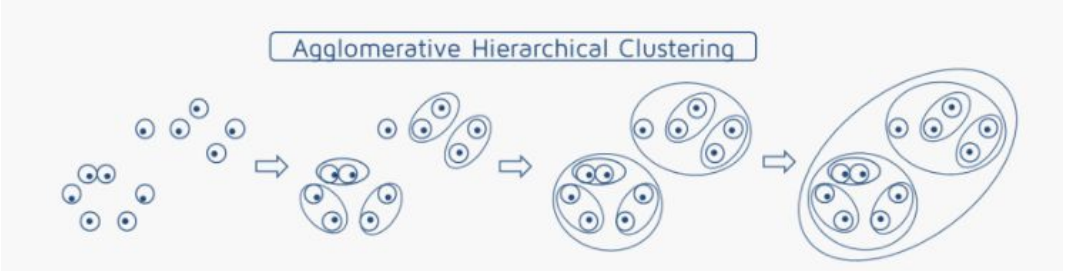
Csoportok képzése:

- input: objektumok az attributumaikkal
- hasonló objektumok meghatározása (cluster)
- unsupervised
- adat redukció
- adat vizualizáció
- hasonlóság / távolság
- határefektus
- zajok (outlier)



HAC klaszterezés

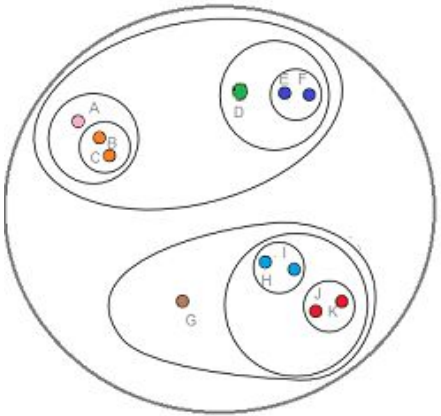
Hierarchikus klaszterező
- induláskor minden objektum egy önálló klaszter
- a legközelebbi klaszterek kerülnek egybevonásra
- megállási feltétel:
  - klaszter darabszám
  - klaszter távolság
- a klaszter darabszám rendszerint előre adott
- nem teljes automatzálás

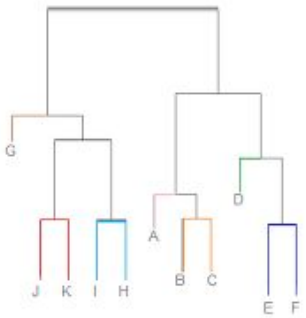


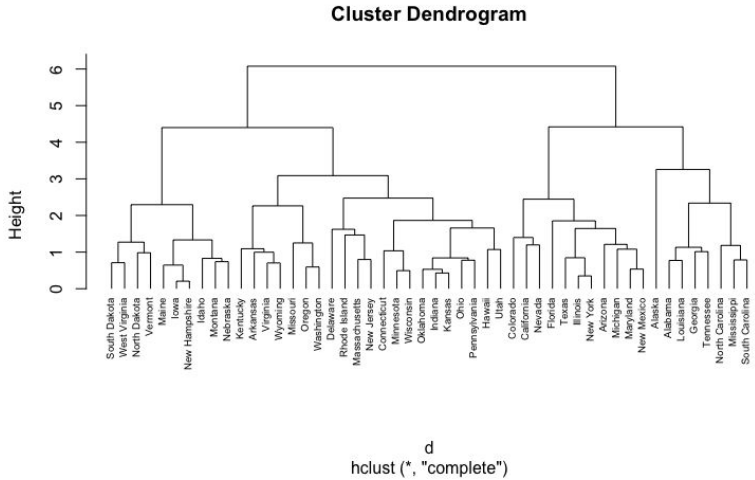
HAC klaszterezés

Dendrogram

- az összefűzési folyamat vizualizációja
- hierarchia







HAC klaszterezés

Cluster távolság

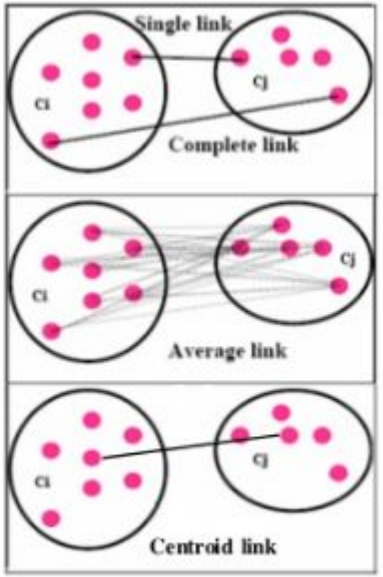
- különböző megközelítések
- centroid
- average
- closest single
- greatest single

Single link (nearest neighbor). The distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters.

Complete link (furthest neighbor). The distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors").

Pair-group average link. The distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters.

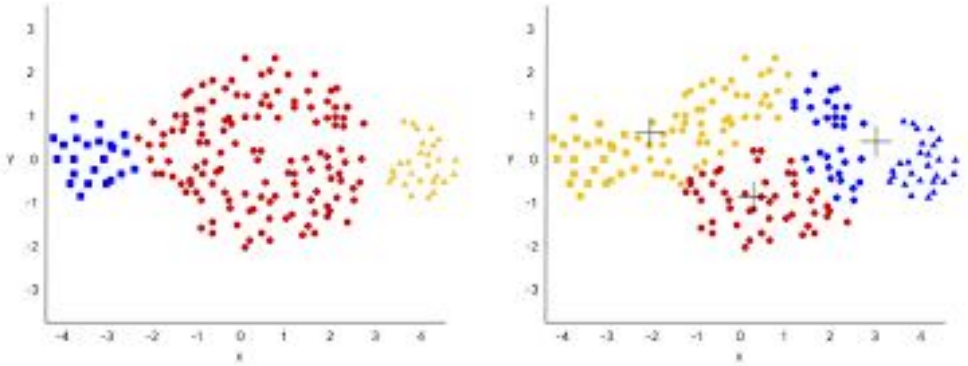
Pair-group centroid. The distance between two clusters is determined as the distance between centroids.

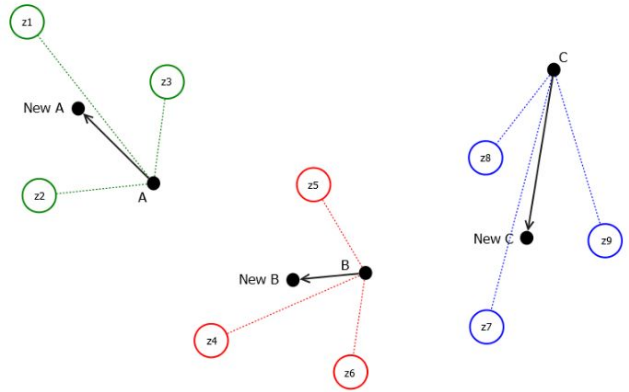


K-means klaszterezés

Klaszterek a centoid-dal adottak
A centroid pozíciók iteratívan javulnak

- klaszterszám megadása
- induló centroid elhelyezés
- tagság meghatározása
- javított centroid pozíció meghatározása
- megállási feltétel: nem javul tovább a centroid pozíció
- konvergencia





K-means klaszterezés

Aktualizálás lépései

- az objektumokat a legközelebbi centroidhoz rendeljük
- a csoportokhoz új centroid kiszámítása
- centroid elmozdulások meghatározása

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set

   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set

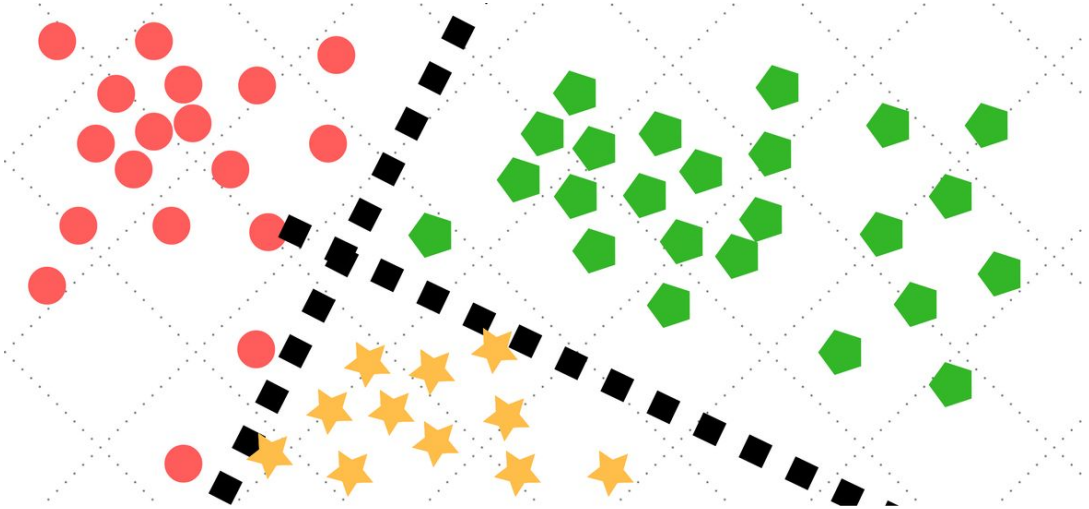
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

Klasszifikáció

Kategória hozzárendelés az objektumokhoz:

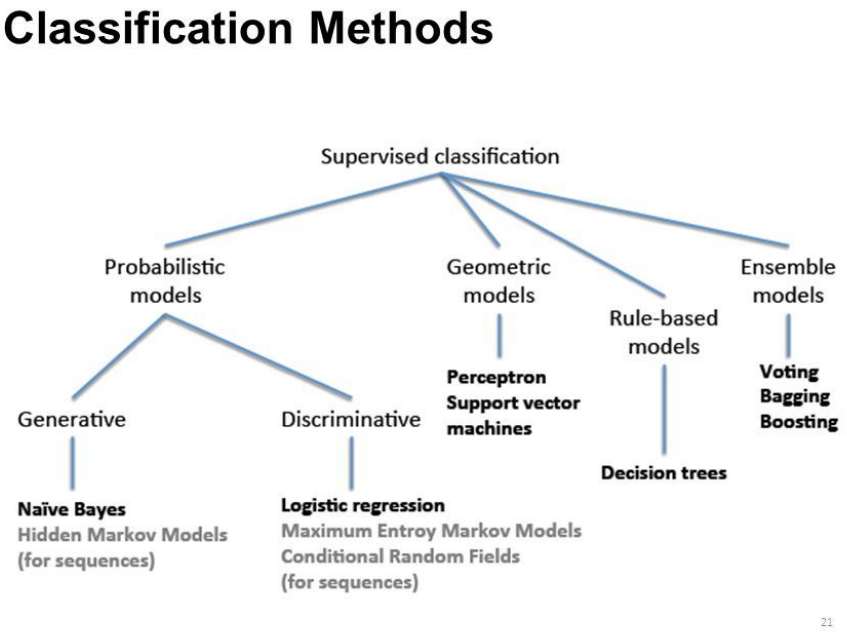
- input: objektumok attributumokkal és kategória értékkel
- előállítja az attribútumok kategóriákra képzését
- felügyelt tanítás (supervised)
- predikciót végezhetünk
- komplex összefüggés kezelése
- multi-class prediction



Osztályozás

Osztályozási módszerek:

- distribution-free
- distribution-based (model)
- probability maximum
- generative
- Bayes-classifier
- rule-based
- Decision-tree
- geometric
- SVM
- neural network
- BPN
- CNN



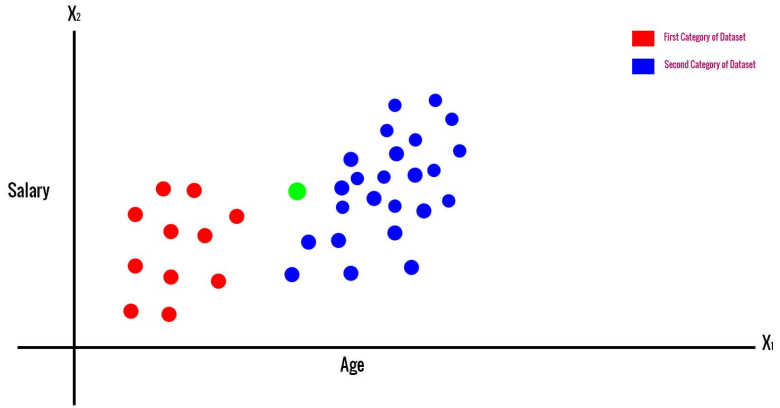
Naive Bayes osztályozó

Módszer jellemzése:

- az egyes kategóriák valószínűségét veszi a attribútumok függvényében
- feltételes valószínűségek számítása
- minden kategória valószínűség számítása
- függetlenségi elv
- egyszerű módszer
- gyors végrehajtás

prediction input: $d(a_1,a_2,....a_m)$
$C = c_1,c_2,..,c_k$

prediction output: c_{win}



Naive Bayes osztályozó

Számítási módszer

- training set: objektumok (attribútum, kategória)
- c|o ritka esemény
- P(c|o) felbontása komponensekre
- P(a|c) kiszámítása
- a|c gyakoribb esemény
- P(c|o) kiszámítása minden c-re

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

$c_w = \text{argmax}_c \{ P(c|d)\}$

$c_w = \text{argmax}_c \{ P(d|c)\ P(c) / P(d)\}$

$c_w = \text{argmax}_c \{ P(d|c)\ P(c) \}$

$c_w = \text{argmax}_c \{ P(a_1\ a_2\ a_3\ ..\ |c)\ P(c) \}$

$c_w = \text{argmax}_c \{ P(a_1\ |c)\ P(\ a_2\ |c)\ P(a_3\ |c)\ ..\ P(c) \}$

Naive Bayes osztályozó

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

		Play Golf	
		Yes	No
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5