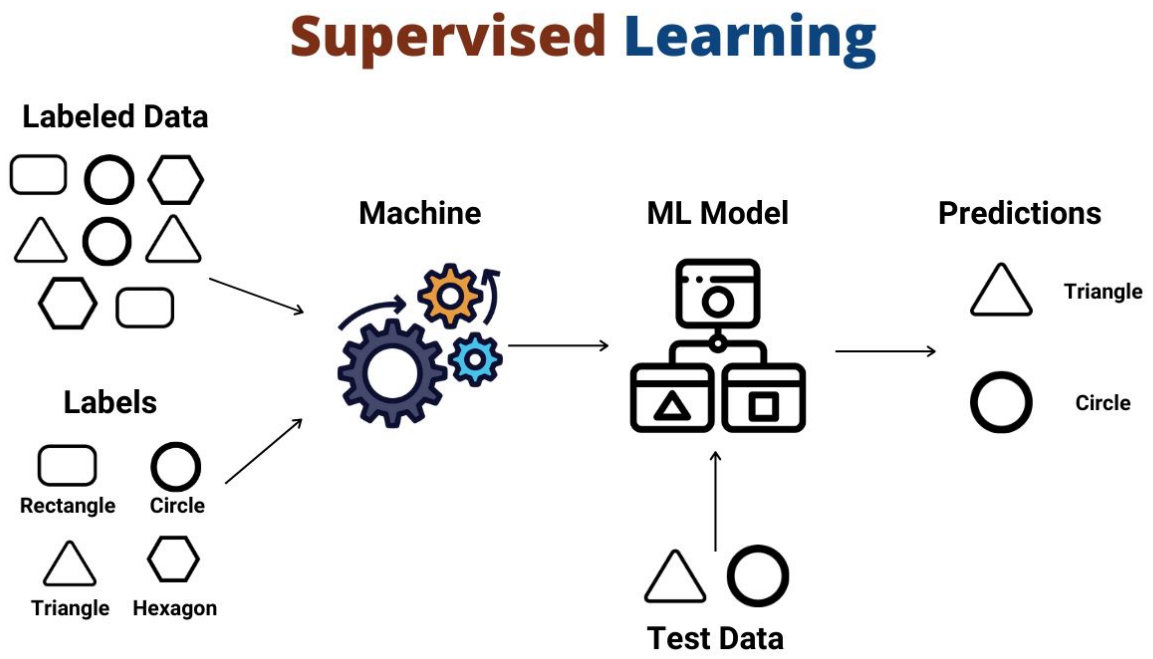
Data Vaders
Projektbemutató

Választott feladat

- Felügyelt gépi tanulási probléma
- Osztályozás
- Adatbeolvasás ⟶ Adatfeldolgozás ⟶ Modellezés ⟶ További fejlesztések
- Megvalósítás az órán használt eszközökkel



Adatbázis

- Banki adathalmaz ügyfélinformációkkal
- Cél: megjósolni, hogy a jövőben mely ügyfelek terveznek megtakarítást elhelyezni a bankban
- Jellemzők: ügyfelek személyes adatai, velük való kapcsolatfelvételekről további információk
- 45211 rekord, 16 jellemző, 1 bináris címke

	age	job	marital	education	default	balance	housing	loan	\
0	58	management	married	tertiary	no	2143	yes	no	
1	44	technician	single	secondary	no	29	yes	no	
2	33	entrepreneur	married	secondary	no	2	yes	yes	
5	35	management	married	tertiary	no	231	yes	no	
6	28	management	single	tertiary	no	447	yes	yes	
...	...	...	...	...	...	...	...	...	
45206	51	technician	married	tertiary	no	825	no	no	
45207	71	retired	divorced	primary	no	1729	no	no	
45208	72	retired	married	secondary	no	5715	no	no	
45209	57	blue-collar	married	secondary	no	668	no	no	
45210	37	entrepreneur	married	secondary	no	2971	no	no	

	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	unknown	5	may	261	1	-1	0	unknown	no
1	unknown	5	may	151	1	-1	0	unknown	no
2	unknown	5	may	76	1	-1	0	unknown	no
5	unknown	5	may	139	1	-1	0	unknown	no
6	unknown	5	may	217	1	-1	0	unknown	no
...	...	...	...	...	...	...	...	...	...
45206	cellular	17	nov	977	3	-1	0	unknown	yes
45207	cellular	17	nov	456	2	-1	0	unknown	yes
45208	cellular	17	nov	1127	5	184	3	success	yes
45209	telephone	17	nov	508	4	-1	0	unknown	no
45210	cellular	17	nov	361	2	188	11	other	no

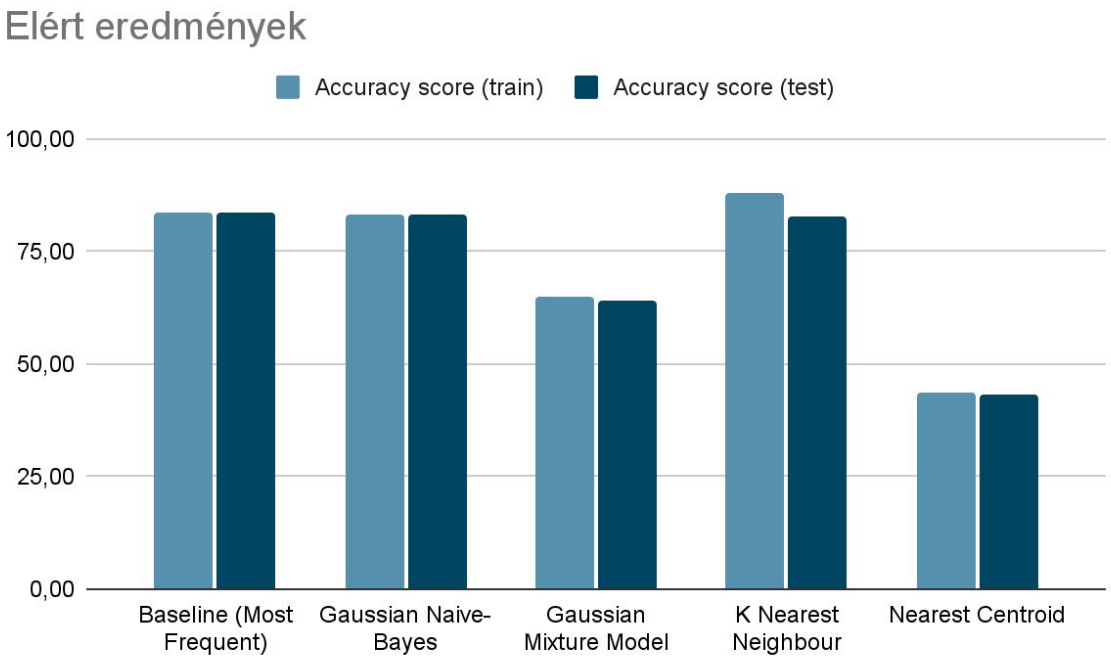
Adatfeldolgozás

- Jellemzők kiválasztása
- Hiányos rekordok törlése
- Szöveges információk számokká alakítása
- Összes jellemző normálása
- Train-dev-test felosztás

	job	marital	education	default	housing	loan	age	balance
0	0.363636	0.5	1.0	0.0	1.0	0.0	1.636763	0.259146
1	0.818182	1.0	0.5	0.0	1.0	0.0	0.305821	-0.436276
2	0.181818	0.5	0.5	0.0	1.0	1.0	-0.739919	-0.445158
5	0.363636	0.5	1.0	0.0	1.0	0.0	-0.549785	-0.369826
6	0.363636	1.0	1.0	0.0	1.0	1.0	-1.215256	-0.298770
...	...	...	...	...	...	...	...	...
45206	0.818182	0.5	1.0	0.0	0.0	0.0	0.971292	-0.174423
45207	0.454545	0.0	0.0	0.0	0.0	0.0	2.872638	0.122957
45208	0.454545	0.5	0.5	0.0	0.0	0.0	2.967705	1.434192
45209	0.090909	0.5	0.5	0.0	0.0	0.0	1.541696	-0.226070
45210	0.181818	0.5	0.5	0.0	0.0	0.0	-0.359650	0.531525

Modellezés

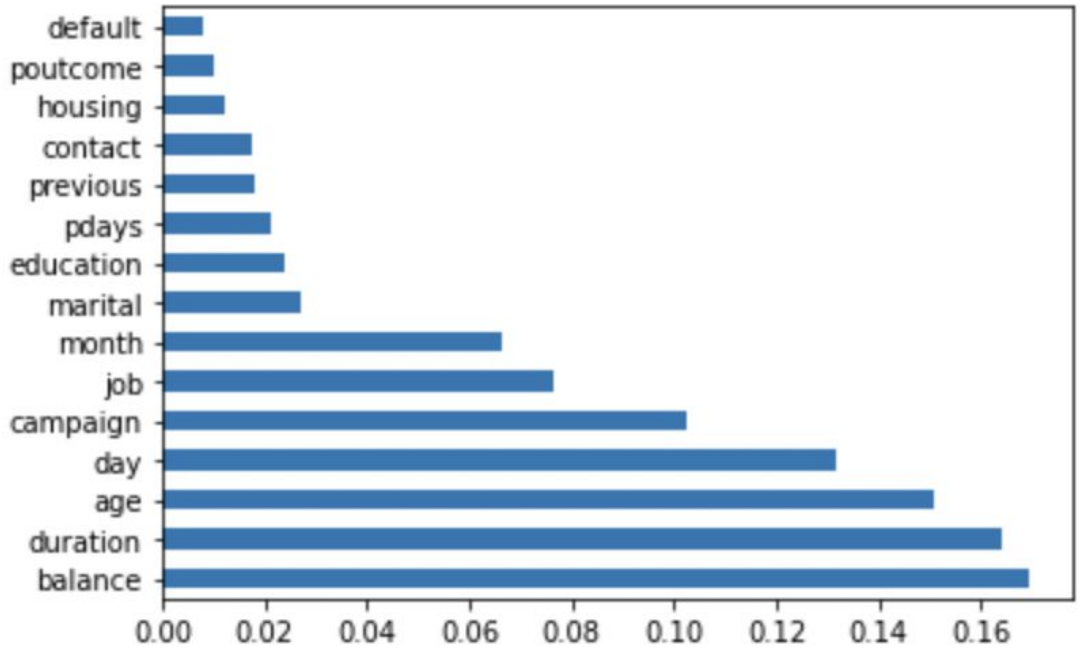
- Kiértékeléshez több különböző metrika
- Egyszerű baseline módszerek
- Gaussian Naive-Bayes
- Gaussian Mixture Model
- K Nearest Neighbour
- Nearest Centroid



További fejlesztések

- Feature selection:
  - VarianceThreshold
  - ExtraTreesClassifier → feature_importances_
- Train-dev-test méretének variálása:
  - Több felosztás kipróbálása
  - Túl- és alultanulás kérdése

	age	day	pdays	previous
40880	-1.025121	-0.458228	-0.412047	-0.250795
34769	0.115686	-1.180563	-0.412047	-0.250795
45194	1.731830	0.023329	1.461755	1.895465
1823	-0.739919	-0.819395	-0.412047	-0.250795
45151	0.591023	-0.819395	3.365458	0.607709
...	...	...	...	...
9793	-0.549785	-0.819395	-0.412047	-0.250795
34720	-0.359650	-1.300952	3.096348	4.470976
15010	0.115686	0.143718	-0.412047	-0.250795
11414	0.971292	0.384496	-0.412047	-0.250795
41428	1.541696	-1.421341	1.162744	1.466213



Köszönjük a figyelmet!