

# Gépitánuulás projekt tervezet – 2018/2019

## 1. Csapat név: kalács123

- a. Tag1: *Kis-Szabó Norbert*
- b. Tag2: *Molnár Ádám*

## 2. Választott feladat

- a. URL: <https://www.kaggle.com/c/petfinder-adoption-prediction>

- b. *Leírás (mi a feladat, feladat típus (osztályozás, regresszió,...), miért ezt választottátok, stb.):*

A feladat kutyák és macskák osztályozása abból a szempontból, hogy milyen gyorsan találhatnak gazdára. Az adathalmaz tartalmaz egyszerű jellemzőket is (kor, nem), nehezebben beolvashatókat (fajta1, fajta2 zajos adatok), valamint képi információkat (képek, videók), így rengeteg jellemző kombináció kipróbálására ad lehetőséget.

- c. *Elérhető adatbázis (formátuma, nyers adat vagy adottak a jellemzők (lehet-e tudni melyik jellemző mit jelent), predikálható érték, stb.):*

A predikálható érték az, hogy milyen gyorsan talál gazdára az adott kisállat. Az adat csv formátumban található meg, az osztályozás szempontjából nem feltétlen releváns, de rengeteg jellemző json vagy kép/videó formátumú segédinformációt tartalmaz. A képek/videók a rekord id-jával, a többi jellemző pedig amelyik label encode-olást kapott dekódolható jsonok segítségével.

train.csv tartalma (kivonat, első 5 elem):

Type	Name	Age	Breed1	Breed2	Gender	Color1	Color2	Color3	MaturitySize	...	Sterilized	Health	Quantity	Fee	State	RescuerID	VideoAmt	PetID	PhotoAmt	AdoptionSpeed	
0	2	Nibble	3	299	0	1	1	7	0	1	...	2	1	1	100	41326	8480853f516546f6cf33aa8cd76c379	0	86e1089a3	1.0	2
1	2	No Name Yet	1	265	0	1	1	2	0	2	...	3	1	1	0	41401	3082c7125d8fb6f7dd4bffa192c8b14	0	6296e909a	2.0	0
2	1	Brisco	1	307	0	1	2	7	0	2	...	2	1	1	0	41326	fa90fa5b1ee11c86938398b60abc32cb	0	3422e4906	7.0	3
3	1	Miko	4	307	0	2	1	2	0	2	...	2	1	1	150	41401	9238e4f44c71a75282e62f7136c6b240	0	5842f1ff5	8.0	2
4	1	Hunter	1	307	0	1	1	0	0	2	...	2	1	1	0	41326	95481e953f8aed9ec3d16fc4509537e8	0	850a43f90	3.0	2

5 rows × 23 columns

jellemzők:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14993 entries, 0 to 14992
Data columns (total 24 columns):
Type                14993 non-null int64
Name                13736 non-null object
Age                14993 non-null int64
Breed1             14993 non-null int64
Breed2             14993 non-null int64
Gender             14993 non-null int64
Color1             14993 non-null int64
Color2             14993 non-null int64
Color3             14993 non-null int64
MaturitySize       14993 non-null int64
FurLength          14993 non-null int64
Vaccinated         14993 non-null int64
Dewormed           14993 non-null int64
Sterilized         14993 non-null int64
Health             14993 non-null int64
Quantity           14993 non-null int64
Fee                14993 non-null int64
State              14993 non-null int64
RescuerID          14993 non-null object
VideoAmt           14993 non-null int64
Description        14981 non-null object
PetID              14993 non-null object
PhotoAmt           14993 non-null float64
AdoptionSpeed      14993 non-null int64
dtypes: float64(1), int64(19), object(4)
```

A jellemzőkről egyértelműen lehet tudni, hogy mi mit jelent.

Néhány fontosabb jellemző bővebben:

1. AdoptionSpeed: meghatározza, hogy a kisállat milyen gyorsan talált gazdára, értékek:
  - 0 - az adatbázisba felvett napon örökbe fogadták
  - 1 - egy hét alatt gazdára talált
  - 2 - 8 és 30 nap között talált gazdát
  - 3 - 31 és 90 nap között talált gazdát
  - 4 - 100 nap után sem talált gazdát (az adatbázis nem tartalmaz ilyen adatokat)
2. Type: meghatározza az állat fajtáját (1 - kutya, 2 - macska)
3. Name: az állat neve (zajos adat)
4. Age: az állat kora
5. Breed: fajtiszta-e az állat, vagy sem
6. Gender: 3 értéket vehet fel:
  - értékek: 1 = hím, 2 = nőstény, 3 = mixed (ilyenkor több állatról van szó, akik egy csoportba tartoznak)
7. MatiritySize: fejlettség méteke (mekkora az adott állat)

értékek: 1 = kistermetű, 2 = közepes, 3 = nagy, 4 = extra nagy, 0 = nem meghatározott

8. FurLength: bunda hossza

értékek: 1 = rövid, 2 = közepes, 3 = hosszú, 0 = nem meghatározott

9. Vaccinated: rendelkezik-e oltásokkal az állat (1 = igen, 2 = nem, 3 = nem biztos)

10. Dewormed: féregtelenített-e (az értékek megegyeznek az előző pontban lévőekkel)

11. Sterilized: ivartalanított-e (1 = igen, 2 = nem, 3 = nem biztos)

12. Health: egészségi állapot

1 = egészséges, 2 = kisebb sérülések, 3 = nagyobb sérülések, 4 = nem behatárolt

13. Fee: az állat ára

d. Nyelv/gépitanuló csomag:

Python nyelvet felhasználva a scikit learn (TPOT), keras és scipy csomagokat használjuk fel.

3. **Baseline rendszer leírása** (jellemzők, gépi tanulási modell, kiértékelés (min és milyen metrikával)):

Teljes adatbázis: 18941

Tanító / kiértékelő adatbázis vágása: 80% / 20%

Kiértékelési metrika: helyesen osztályozott egyedek aránya (accuracy)

Baseline rendszer:

GMM osztályozó, előre adott jellemzőkön  
pontosság: 28.67%

#### 4. Feladat felosztás:

##### a. Jellemzőkinyerés: *Kis-Szabó Norbert*

###### i. Tervezett feladatok:

1. adatok átalakítása típusuk szerint (nominális, intervallum, ordinális)
2. dimenziócsökkentés alkalmazása korrelált adatokra (ilyenek az egészségügyi adatok)

##### b. Géptanuló modellek összehasonlítása: *Molnár Ádám*

###### i. Tervezett feladatok:

1. Következő modellek kipróbálása:
  - a. Gaussian Mixture Model
  - b. RandomForestClassifier
  - c. DecisionTreeClassifier
  - d. Alacsonyabb rétegszámú (3) neuronháló
  - e. Nearest Neighbors
2. GridSearch paraméter optimalizálás a fenti modellekre
3. A fent modellek hatékonyságának összehasonlítása

##### c. További technikák:

###### i. Tag1: *Kis-Szabó Norbert*

1. annak a vizsgálata, hogy a hanyagul felvett adatok hogyan korrelálnak az adaptáció sebességével (leírás jellemző vizsgálata)

###### ii. Tag2: *Molnár Ádám*

1. TPOT segítségével megtalálni a legjobb modellt
2. A legjobban teljesítő modellek összehasonlítása más metrikák szerint

##### d. Eredmények:

###### i. Tag1: *Kis-Szabó Norbert*

1. adatok zajosságának csökkentése
2. dimenziócsökkentő eljárások vizsgálata

###### ii. Tag2: *Molnár Ádám*

1. Osztályozók hatékonyságainak összehasonlítása egymással, és baseline eredményekkel, és ezek grafikus megjelenítése
2. Modellek hibáinak elemzése