1η Σειρά ασκήσεων - Report

Κωνσταντίνος Κυριάκου 3015
Ομάδα 93

Γενικές σημειώσεις υλοποίησης.

Η υλοποίηση πραγματοποιήθηκε εξ' ολοκλήρου σε μορφή notebook του Kaggle και το αρχείο κώδικα είναι τύπου ".ipynb". Χρησιμοποιήθηκαν οι βιβλιοθήκες της Sklearn για τις μεθόδους kNN, SVM και Tensorflow Keras για τα νευρωνικά δίκτυα.

Κατανόηση δεδομένων

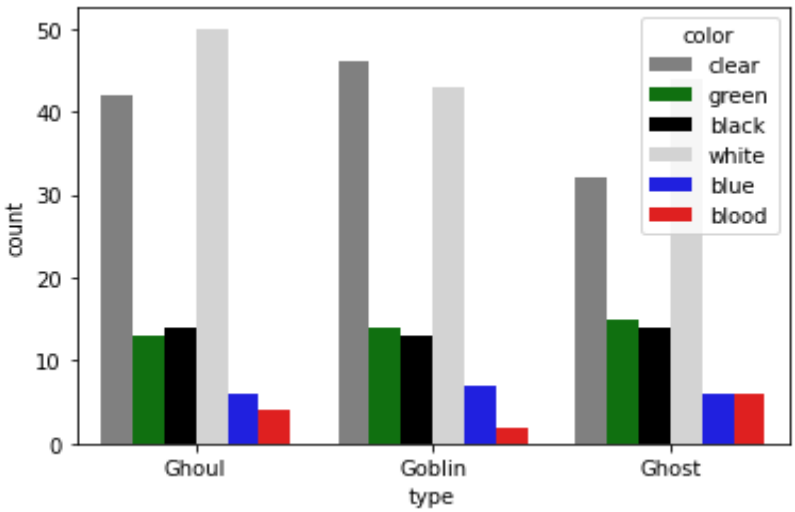
Το πρώτο βήμα που πρέπει να κάνουμε είναι να κατανοήσουμε τα δεδομένα που μας δίνονται για εκπαίδευση. Πιο συγκεκριμένα, πρέπει να εντοπίσουμε ποιά είναι χαρακτηριστικά που μας βοηθούν να αποφασίσουμε τον τύπο ενός τέρατος.

```
train_df
```

	id	bone_length	rotting_flesh	hair_length	has_soul	color	type
0	0	0.354512	0.350839	0.465761	0.781142	clear	Ghoul
1	1	0.575560	0.425868	0.531401	0.439899	green	Goblin
2	2	0.467875	0.354330	0.811616	0.791225	black	Ghoul
3	4	0.776652	0.508723	0.636766	0.884464	black	Ghoul
4	5	0.566117	0.875862	0.418594	0.636438	green	Ghost
...	...	...	...	...	...	...	...
366	886	0.458132	0.391760	0.660590	0.635689	blue	Goblin
367	889	0.331936	0.564836	0.539216	0.551471	green	Ghost
368	890	0.481640	0.501147	0.496446	0.544003	clear	Ghoul
369	896	0.294943	0.771286	0.583503	0.300618	clear	Ghost
370	897	0.670200	0.768469	0.737274	0.608384	white	Ghoul

371 rows × 7 columns

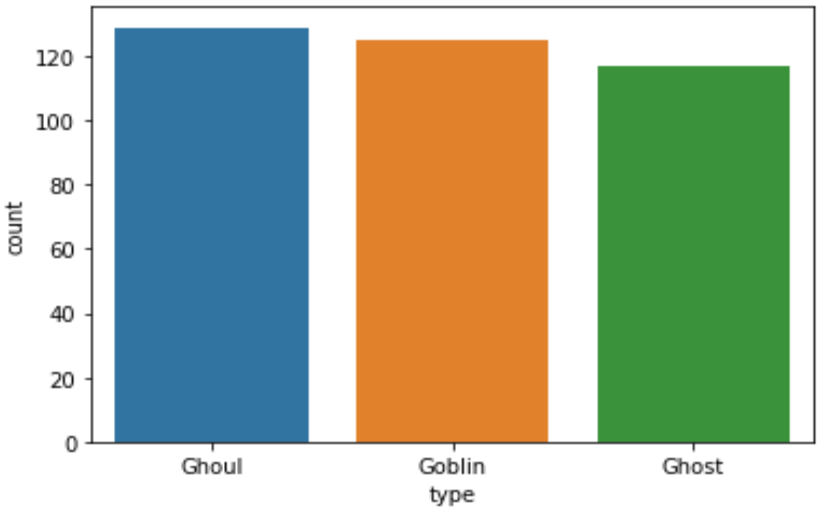
Αντίστοιχα, πρέπει να εντοπιστούν ποια χαρακτηριστικά δεν συμβάλλουν κάτι στην διαδικασία απόφασης. Στην υλοποίηση δεν λήφθηκε σε καμία μέθοδο υπόψην το χρώμα του κάθε τέρατος.



Παρατηρούμε ότι ανάμεσα στις 3 κατηγορίες δεν υπάρχει κάποια αξιοσημείωτη εξάρτηση ανάμεσα σε κατηγορία και χρώμα.

Για προφανείς λόγους το χαρακτηριστικό 'id ' θα εξαιρεθεί από την εκπαίδευση παρομοίως.

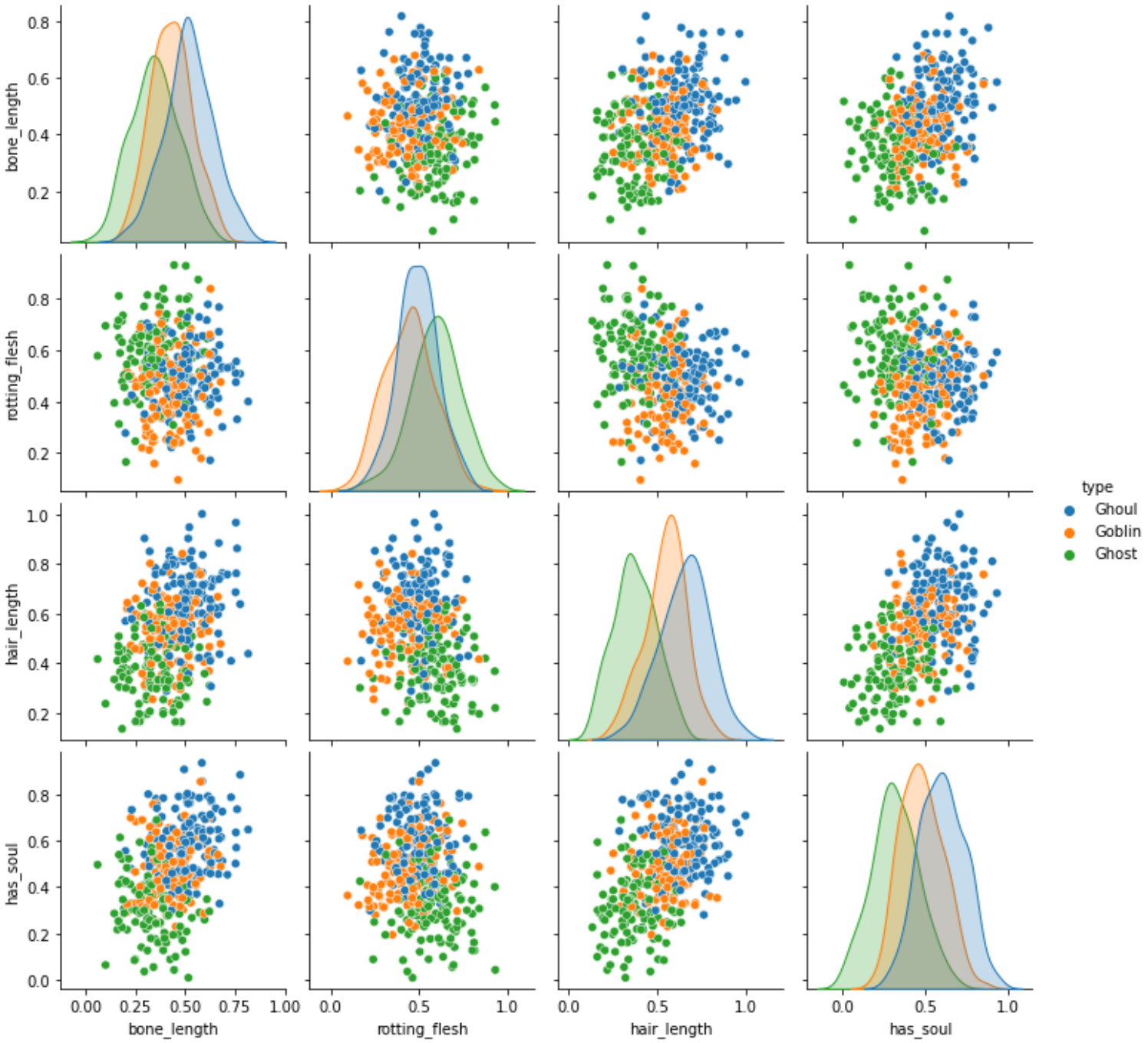
Επιπροσθέτως πειραματικά απεικονίστηκε το πλήθος των τεράτων κάθε κατηγορίας για να δούμε αν κάποια κατηγορία υπερτερεί έναντι των άλλων.



Δεν υπερτερεί κάποια κατηγορία, ωστόσο η "Ghoul" είναι η συχνότερη και η "Ghost" η λιγότερο συχνή.

Τελικά τα χαρακτηριστικά που θα αξιοποιήσουν οι μέθοδοί μας για να αποφασίσουν είναι οι *“bone_length”, “rotting_flesh”, “hair_length”, “has_soul”*.

Παρακάτω απεικονίζονται οι τιμές των χαρακτηριστικών ανά κατηγορία. Έτσι καταλαβαίνουμε καλύτερα σε τι χαρακτηριστικά θα τείνει η κάθε κατηγορία τέρατος.



Π.χ. Παρατηρείται ότι τα τέρατα “Goblin” έχουν κατά μέσο όρο μεγαλύτερο “hair_length” απ’ ότι οι άλλες κατηγορίες.

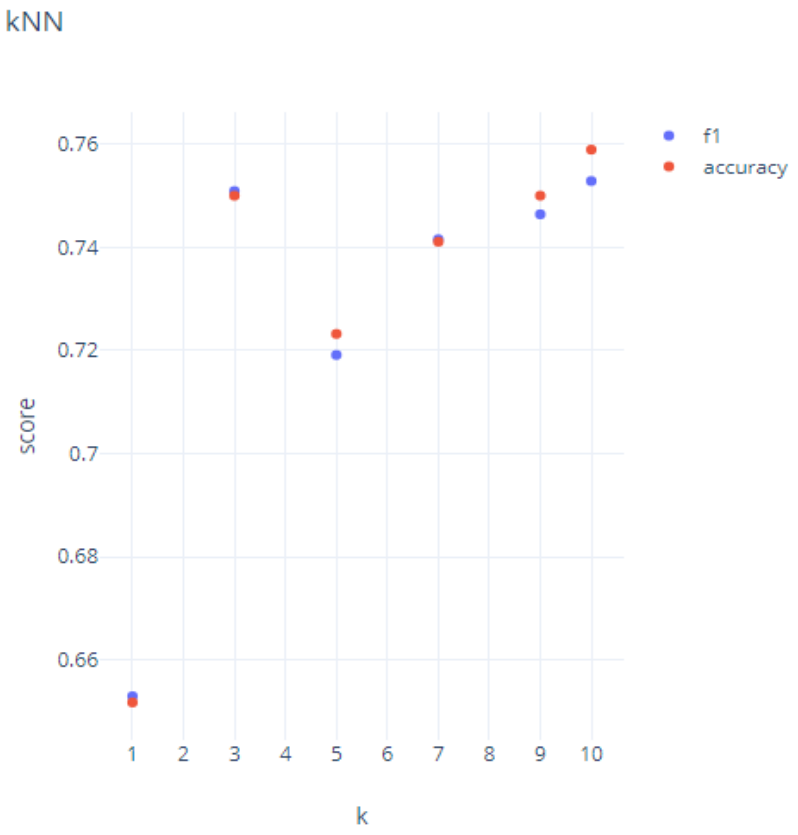
Εκτέλεση και αποτελέσματα

kNN

Στην μέθοδο kNN η απόδοσή της καθορίζεται από τον αριθμό των γειτόνων “k”. Πειραματικά, εκτελέστηκε η μέθοδος για k=1 , 3, 5, 10, 11 απ’ όπου προέκυψαν τα παρακάτω scores στην εκπαιδευση (σε όλα τα αποτελέσματα έγινε στρογγυλοποίηση στα 4 σημαντικά ψηφία) :

k	Accuracy	F1
1	0.6518	0.6529
3	0.75	0.7509
5	0.7232	0.7191
7	0.7411	0.7415
9	0.75	0.7464
10	0.7589	0.7528
11	0.7768	0.7724

Στον τελικό κώδικα που παραδόθηκε, η μέθοδος εκτελείται με k=11 διότι εκεί προέκυψε το μέγιστο accuracy.



Νευρωνικό δίκτυο #1

Παρακάτω βλέπουμε πως επηρεάζονται τα scores που πετυχαίνει το δίκτυο από την μεταβολή του αριθμού των νευρώνων στο κρυμμένο επίπεδο του (όλα τα πειράματα εκτελέστηκαν με αριθμό εποχών = 40) :

K	Accuracy	F1
50:	0.3125	0.1489
100	0.3214	0.1564
200	0.3214	0.2102

Στο τελικό πρόγραμμα επιλέχθηκε K = 200

Νευρωνικό δίκτυο #2

Για τους συνδυασμούς που αναγράφονται έχουμε τα scores:

(K1, K2)	Accuracy	F1
(50, 25)	0.1875	0.1349
(100, 50)	0.3214	0.1564
(200, 100)	0.3125	0.1488

Στο τελικό πρόγραμμα επιλέχθηκε (K1,K2) = (200, 100)

SVM

Τέλος, για τις μεθόδους SVM:

	Linear	Gaussian (Gamma: Scale)	Gaussian (Gamma: Auto)
accuracy:	0.7321	0.75	0.7321
F1:	0.7255	0.75231	0.7302

Επιλογή καλύτερης μεθόδου

Η επιλογή της μεθόδου προς υποβολή στο Kaggle γίνεται βάσει του accuracy score που υπολογίστηκε από το σύνολο εκπαίδευσης. Στην συγκεκριμένη έκδοση επιλέγεται η kNN με k=11.

Το τελικό accuracy score που υπολογίζεται από το Kaggle γι αυτή την μέθοδο είναι 0.72967.

Προβληματισμοί - Βελτιώσεις

- Θα μπορούσαν να χρησιμοποιηθούν επιπροσθέτως συνδυαστικές τιμές για τα χαρακτηριστικά, π.χ. “rotting_flesh + hair_length”.
- Τα νευρωνικά δίκτυα που δημιουργήθηκαν δεν φαίνεται να κάνουν έγκυρες προβλέψεις, ειδικά σε σύγκριση με τις άλλες μεθόδους. Πιο πολύπλοκα και πολυεπίπεδα νευρωνικά δίκτυα ίσως λύνουν το πρόβλημα αυτό.
- Για την παραμετροποίηση όλων των μεθόδων θα μπορούσε να χρησιμοποιηθεί η βιβλιοθήκη της Sklearn GridSearchCV, απ’ όπου θα κρατούσαμε την τιμή με το βέλτιστο accuracy αυτόματα.