

ZOO ANIMALS CLASSIFICATION

MIC MACHINE LEARNING ASSIGNMENT

Victoria Trinita Pardede



OBJEKTIF

Mengklasifikasikan binatang yang ada di kebun binatang berdasarkan ciri-cirinya.

Binatang akan diklasifikasikan ke dalam kelas-kelas berikut:

Kelas 1 - Mamalia

Kelas 2 - Unggas

Kelas 3 - Amphibi

Kelas 4 - Ikan

Kelas 5 - Katak

Kelas 6 - Serangga

Kelas 7 - Lainnya

ANALISIS DATASET

zoo.data
101 instances
18 features

target feature



Independent Feature				Dependent Feature
animal_name (unique for each instance))	airborne (boolean)	breathes (boolean)	tail (boolean)	Type (numeric integer values in range [1,7])
hair (boolean)	aquatic (boolean)	venomous (boolean)	domestic (boolean)	
feathers (boolean)	predator (boolean)	fins (boolean)	catsize (boolean)	
eggs (boolean)	toothed (boolean)	legs (numeric set of values: {0,2,4,5,6,8})		
milk (boolean)	backbone (boolean)			

ALGORITME MACHINE LEARNING

Pada pemodelan, digunakan algoritme:

1. Decision Tree

Memecahkan masalah menggunakan representasi pohon. Setiap node internal pohon merepresentasikan atribut, dan setiap node leaf merepresentasikan label kelas.

Pseudocode:

1. Tempatkan atribut terbaik dari dataset di root pohon.
2. Bagi set training menjadi beberapa subset. Subset harus dibuat sedemikian rupa agar setiap subset berisi data dengan nilai yang sama untuk suatu atribut.
3. Ulangi langkah 1 dan langkah 2 pada setiap subset sampai node leaf di semua cabang ditemukan.

ALGORITME MACHINE LEARNING

2. Random Forest

Algoritma klasifikasi yang terdiri dari banyak decision tree. Random Forest menggunakan bagging dan feature randomness saat membuat setiap pohon agar terbentuk 'hutan' yang berisi pohon yang tidak berkorelasi, sehingga hasil prediksinya lebih akurat dibanding hasil prediksi satu pohon saja. Keputusan akhir dibuat berdasarkan mayoritas tree dan dipilih oleh random forest.

Pseudocode:

1. Pilih secara acak "k" fitur dari total "m" fitur.
2. Di mana $k \ll m$
Di antara "k" fitur tadi, hitung node "d" menggunakan split point terbaik.
3. Bagi node tersebut menjadi daughter nodes menggunakan split terbaik.
4. Ulangi langkah 1-3 hingga tercapai "l" jumlah node.
5. Bangun 'hutan' dengan mengulangi langkah 1-4 sebanyak "n" kali untuk membuat "n" pohon.

ALGORITME MACHINE LEARNING

3. K-Nearest Neighbors

Algoritma klasifikasi yang terdiri dari banyak decision tree. Random Forest menggunakan bagging dan feature randomness saat membuat setiap pohon agar terbentuk 'hutan' yang berisi pohon yang tidak berkorelasi, sehingga hasil prediksinya lebih akurat dibanding hasil prediksi satu pohon saja. Keputusan akhir dibuat berdasarkan mayoritas tree dan dipilih oleh random forest.

Pseudocode:

1. Muat data training dan data tes.
2. Tentukan nilai 'k'
3. Untuk setiap titik dalam data tes:
 - temukan jarak Euclidean ke semua titik data training
 - simpan jarak Euclidean dalam list dan urutkan
 - pilih 'k' poin pertama
 - tugaskan sebuah class ke titik tes berdasarkan mayoritas kelas yang ada di titik yang dipilih
4. Selesai



METODE EVALUASI

1. Cross Validation

Persentase klasifikasi yang benar diperoleh melalui metode cross validation. Biasanya dilakukan pada dataset training. Merupakan metode model fine-tuning dengan me-resampling dataset ke dalam training dan validasi sebanyak 'k' kali.

2. Classification Report

Digunakan untuk mengukur kualitas prediksi dari algoritma klasifikasi. True Positive, False Positive, True negative dan False Negatives digunakan untuk memprediksi metrik laporan klasifikasi

3. Confusion Matrix

Matriks (tabel) yang dapat digunakan untuk mengukur kinerja algoritma machine learning, biasanya supervised learning. Setiap baris mewakili instance dari kelas asli dan setiap kolom mewakili instance dari kelas yang diprediksi.



HASIL DECISION TREE

Cross Validation

```
↳ Accuracy: 0.96 (+/- 0.13)  
/usr/local/lib/python3.6/dist-packages/sklearn/model_selection/_split.py:667:  
% (min_groups, self.n_splits)), UserWarning)
```

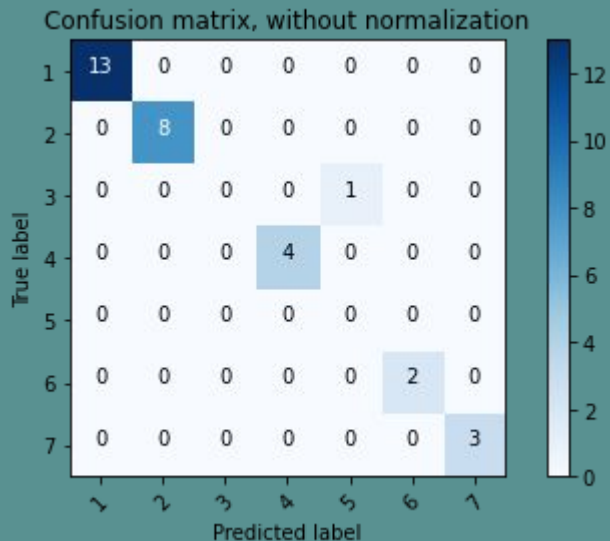

HASIL DECISION TREE

Classification Report

Classification Report:					
	precision	recall	f1-score	support	
1	1.00	1.00	1.00	13	
2	1.00	1.00	1.00	8	
3	0.00	0.00	0.00	1	
4	1.00	1.00	1.00	4	
5	0.00	0.00	0.00	0	
6	1.00	1.00	1.00	2	
7	1.00	1.00	1.00	3	
accuracy			0.97	31	
macro avg	0.71	0.71	0.71	31	
weighted avg	0.97	0.97	0.97	31	

HASIL DECISION TREE

Confusion Matrix



Contoh membaca:

Terdapat 13 binatang yang diprediksi merupakan anggota kelas 1, dan memang benar (true label).



HASIL RANDOM FOREST

Cross Validation

```
↳ /usr/local/lib/python3.6/dist  
   % (min_groups, self.n_split  
   Accuracy: 0.97 (+/- 0.09)
```

HASIL RANDOM FOREST

Classification Report

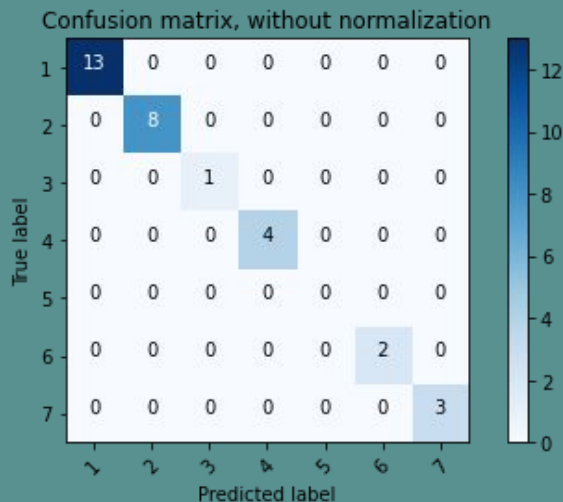
```
Classification Report:
              precision    recall  f1-score   support

     1         1.00      1.00      1.00        13
     2         1.00      1.00      1.00         8
     3         1.00      1.00      1.00         1
     4         1.00      1.00      1.00         4
     6         1.00      1.00      1.00         2
     7         1.00      1.00      1.00         3

 accuracy          1.00          1.00      1.00        31
 macro avg         1.00          1.00      1.00        31
 weighted avg      1.00          1.00      1.00        31
```

HASIL RANDOM FOREST

Confusion Matrix



Contoh membaca:

Terdapat 13 binatang yang diprediksi merupakan anggota kelas 1, dan memang benar (true label).



HASIL K-NEAREST NEIGHBORS

Cross Validation

```
↳ Accuracy: 0.95 (+/- 0.10)  
/usr/local/lib/python3.6/dist-pa  
% (min_groups, self.n_splits)
```

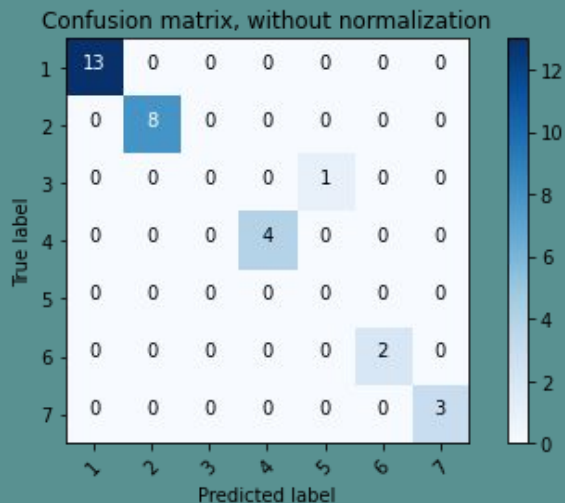
HASIL K-NEAREST NEIGHBORS

Classification Report

Classification Report:				
	precision	recall	f1-score	support
1	1.00	1.00	1.00	13
2	1.00	1.00	1.00	8
3	0.00	0.00	0.00	1
4	1.00	1.00	1.00	4
5	0.00	0.00	0.00	0
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	3
accuracy			0.97	31
macro avg	0.71	0.71	0.71	31
weighted avg	0.97	0.97	0.97	31

HASIL K-NEAREST NEIGHBORS

Confusion Matrix



Contoh membaca:

Terdapat 13 binatang yang diprediksi merupakan anggota kelas 1, dan memang benar (true label).