Wstęp do analizy danych - raport

AGNIESZKA TRACZ, PIOTR LACHOWICZ, PIOTR ZAJĄC

CZERWIEC 2023

Opis zbioru danych

- Źródło: nasze dane pobraliśmy ze strony https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data
- Sposób gromadzenia: dane zbierane były przez badaczy. Monitorowali oni trzy gniazda pingwinów, znajdujące się na trzech różnych wyspach. Do zbierania danych zastosowano metody terenowe i specjalistyczne narzędzia.
- Zbiór danych składa się z 344 obserwacji dotyczących trzech gatunków pingwinów: Adelie, Chinstrap i Gentoo. Dane zawierają m. in. nazwę gatunku, wymiary dzioba, wagę pingwina, wyspę, z której pochodzi oraz płeć. Są 3 cech jakościowe (w tym gatunek) i 4 ilościowe.

Temat i cel analizy

Celem naszej analizy będzie sklasyfikowanie, do którego z trzech gatunków należy pingwin.

Metody użyte w projekcie:

- drzewo decyzyjne
- KNN
- naiwna klasyfikacja Bayesa

Potencjalne zastosowanie modelu: model może być pomocny w pracy badaczy, ornitologów etc.

Eksploracja i przygotowanie danych

Braki w danych: W zbiorze danych brakowało w sumie 18 wartości. Po eksploracji okazało się, że 10 wierszy zawiera NA. Wszystkie te wiersze zostały usunięte. Oprócz tego usunięto też wiersz, w którym zmienna sex przyjmowała wartość „.”, nie mającą większego sensu i będącą prawdopodobnie pomyłką.

```
> summary(penguins)
      species          island    culmen_length_mm culmen_depth_mm flipper_length_mm  body_mass_g       sex
 Adelie   :152   Biscoe   :168   Min.   :32.10    Min.   :13.10   Min.   :172.0     Min.   :2700   .      :  1
 Chinstrap: 68   Dream    :124   1st Qu.:39.23    1st Qu.:15.60   1st Qu.:190.0     1st Qu.:3550   FEMALE:165
 Gentoo   :124   Torgersen: 52   Median :44.45    Median :17.30   Median :197.0     Median :4050   MALE   :168
                                 Mean   :43.92    Mean   :17.15   Mean   :200.9     Mean   :4202   NA's   : 10
                                 3rd Qu.:48.50    3rd Qu.:18.70   3rd Qu.:213.0     3rd Qu.:4750
                                 Max.   :59.60    Max.   :21.50   Max.   :231.0     Max.   :6300
                                 NA's   :2        NA's   :2       NA's   :2         NA's   :2
```

Po tej modyfikacji w zbiorze danych pozostał 333 obserwacji — wystarczająco duży zbiór do przeprowadzenia analizy.

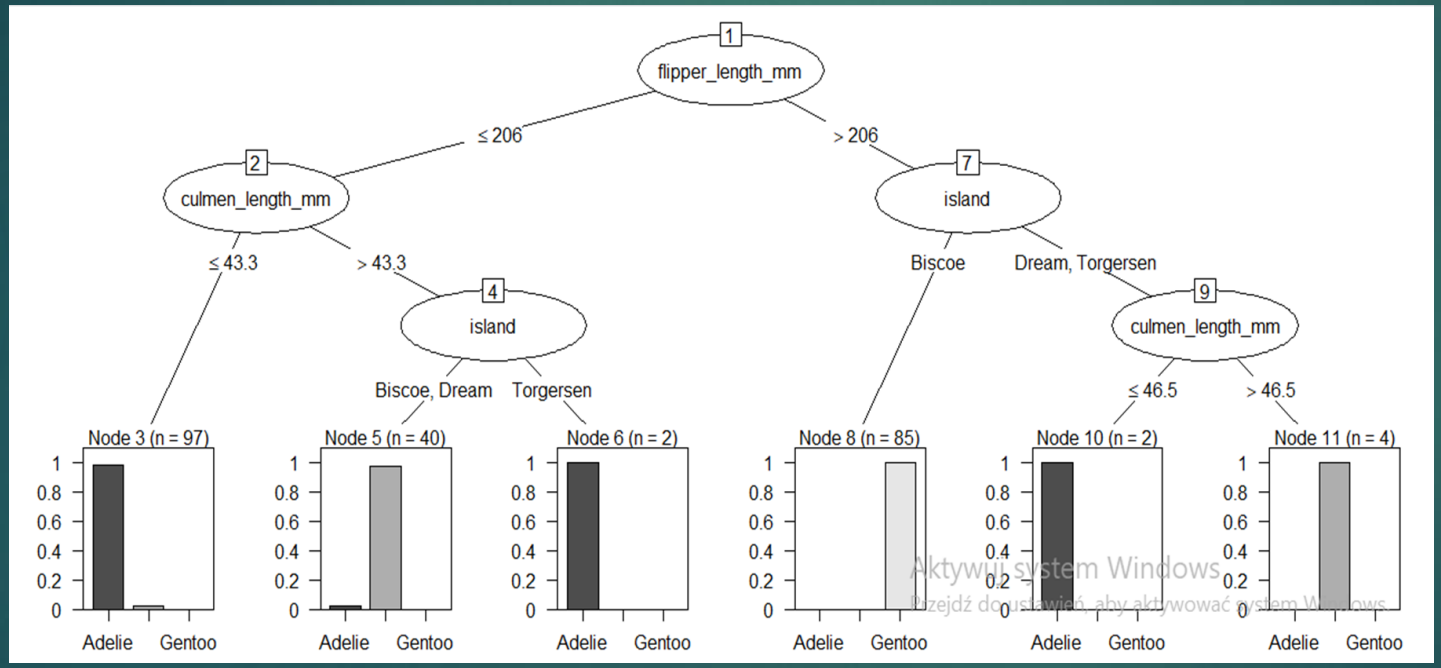
Zbiór testowy i uczący

Nasze dane podzieliśmy na dane uczące i testowe w proporcjach odpowiednio: 70% i 30%. W każdym z trzech modeli używamy tego samego zbioru uczącego i testowego.

Następnie sprawdziliśmy jakie są proporcje klas w każdym z tych zbiorów i porównaliśmy je do proporcji między klasami z wyjściowego zbioru — okazały się być zbliżone.

Model 1 Drzewo decyzyjne

Zastosowanie tego modelu nie wymagało żadnych modyfikacji danych. Zdecydowaliśmy się na jego zastosowanie ze względu na czytelną interpretacje jaką dają drzewa (rys. poniżej).



Ocena modelu

- Nasz model poprawnie sklasyfikował 98 pingwinów, co daje skuteczność na poziomie około 95 %, zatem satysfakcjonująco wysoką.

	predicted default			
actual default	Adelie	Chinstrap	Gentoo	Row Total
Adelie	45 0.437	1 0.010	0 0.000	46
Chinstrap	3 0.029	20 0.194	0 0.000	23
Gentoo	0 0.000	1 0.010	33 0.320	34
Column Total	48	22	33	103

Poprawa modelu za pomocą AdaBoost

Postanowiliśmy poprawić poprzedni model, używając boostingu z parametrem

trials = 10 (algorytm zbudował 10 drzew decyzyjnych, za pomocą których jest dokonywana klasyfikacja).

Skuteczość uległa poprawie i wyniosła około 99%. Przez to, że model jest bardziej skomplikowany, nie jest już tak łatwy w interpretacji jak pojedyncze drzewo.

	predicted default			
actual default	Adelie	Chinstrap	Gentoo	Row Total
Adelie	46 0.447	0 0.000	0 0.000	46
Chinstrap	1 0.010	22 0.214	0 0.000	23
Gentoo	0 0.000	0 0.000	34 0.330	34
Column Total	47	22	34	103

Model 2:
Naiwny klasyfikator Bayesowski

	predicted default			
actual default	Adelie	Chinstrap	Gentoo	Row Total
Adelie	46	0	0	46
	0.447	0.000	0.000	
Chinstrap	0	23	0	23
	0.000	0.223	0.000	
Gentoo	0	0	34	34
	0.000	0.000	0.330	
Column Total	46	23	34	103

Nasz model ma 100% skuteczność. Poprawnie sklasyfikował wszystkie gatunki pingwinów.

Naiwny klasyfikator Bayesowski lepiej poradził sobie z problemem klasyfikacji niż drzewo decyzyjne i AdaBoost.

Model 3: Metoda K-NN

- Do tego modelu musieliśmy specjalnie przygotować nasze dane.
- Nasze modyfikacje polegały na zakodowaniu danych jakościowych na ilościowe (za pomocą dummy-coding). Dotyczy to kolumn: płeć i wyspa.
- Następnie znormalizowaliśmy nasze dane, ponieważ tego wymaga nasza metoda.

Wnioski z KNN:

Dla k należącego do zbioru {3,5,7,8,9,10,11,12,13,14,15,18,19,20,22,23,24} dokładność wynosi 99.02913%.

	predicted default			
actual default	Adelie	Chinstrap	Gentoo	Row Total
Adelie	46 0.447	0 0.000	0 0.000	46
Chinstrap	1 0.010	22 0.214	0 0.000	23
Gentoo	0 0.000	0 0.000	34 0.330	34
Column Total	47	22	34	103

Dla k= 1 lub k=2 dokładność wyniosła 100%

	predicted default			
actual default	Adelie	Chinstrap	Gentoo	Row Total
Adelie	46 0.447	0 0.000	0 0.000	46
Chinstrap	0 0.000	23 0.223	0 0.000	23
Gentoo	0 0.000	0 0.000	34 0.330	34
Column Total	46	23	34	103

Dla k powyżej 24 dokładność jest gorsza

Podsumowanie:

Udało nam się zbudować skuteczny model do klasyfikacji gatunków pingwinów.

Największą skuteczność miał model naiwnego klasyfikatora Bayesowskiego,

aż 100% dokładności. Za raz po nim jest Model KNN.

Najgorzej z tym problemem poradził sobie model drzewa decyzyjnego,

Natomiast był on dla nas najbardziej czytelny.