



*Studio delle specie di
pinguini presenti
nell'arcipelago di
Palmer*

Alberto Biliotti

Matricola: 7109894

alberto.biliotti@stud.unifi.it

Introduzione

In questo lavoro mi sono occupato dello studio e della classificazione delle tre specie di pinguino presenti nell'arcipelago di Palmer, situato al largo della costa nord-occidentale della penisola antartica. Queste tre specie sono: il Pigoscelide Antartico, il Pinguino di Adelia e il Pigoscelide comune (nel dataset vengono indicate rispettivamente come «Chinstrap», «Adelie» e «Gentoo»). I pinguini in oggetto provengono in particolare da tre isole dell'arcipelago: Dream, Torgersen e Biscoe.



ESEMPLARE DI PIGOSCELIDE COMUNE « GENTOO »

Chinstrap



Gentoo



Adélie



Data understanding

Per prima cosa è importante comprendere le caratteristiche del dataset su cui andremo a lavorare, per poter osservare eventuali pattern o correlazioni tra i dati

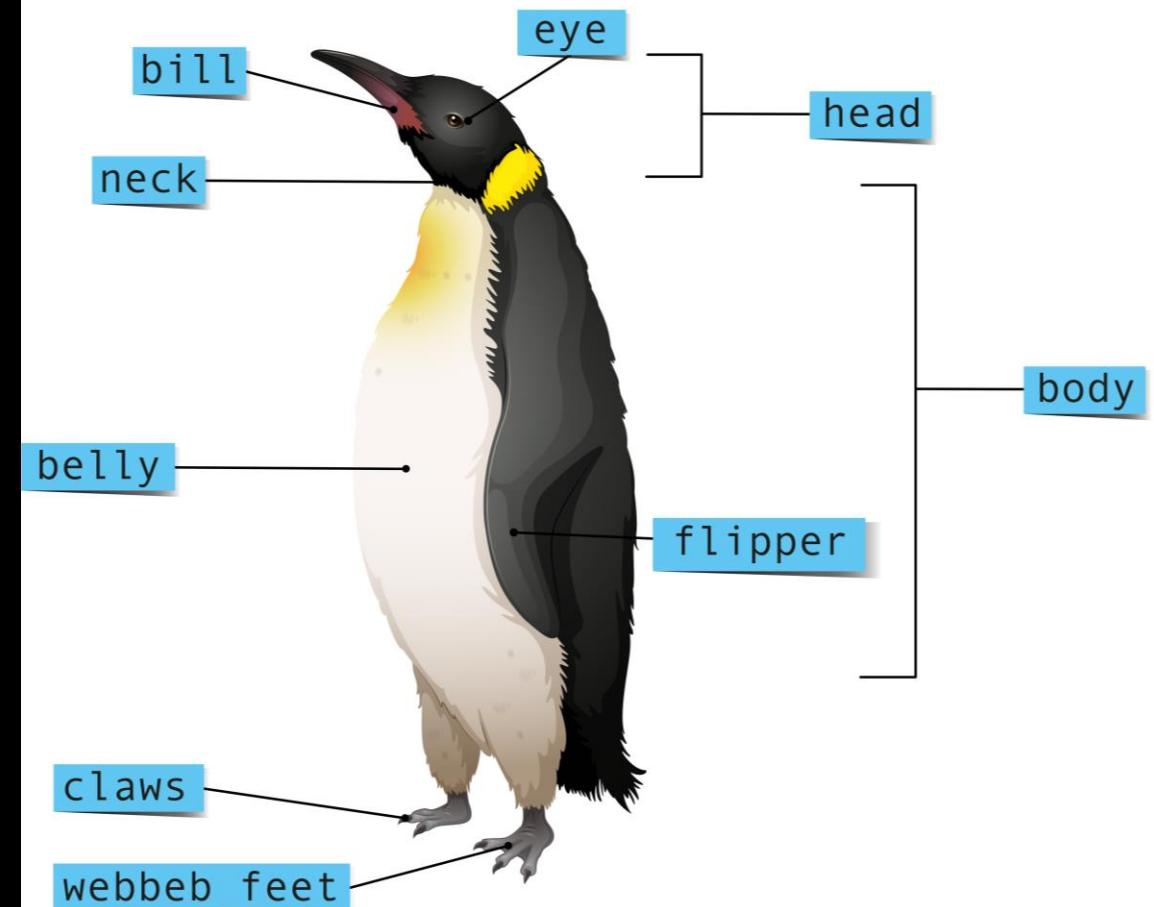
Caratteristiche del dataset

Nel dataset preso in esame, per ogni pinguino sono riportati sette attributi di cui tre categorici: l'isola di provenienza (Dream, Torgersen, o Biscoe), il genere e la specie di appartenenza (Gentoo, Adelie, Chinstrap), oltre a quattro attributi continui: la massa corporea in grammi (body_mass_g), la lunghezza e la profondità del becco (bill_length_mm e bill_depth_mm) e la lunghezza delle ali (flipper_length_mm) in millimetri.

Il dataset riporta le osservazioni effettuate su 344 esemplari di pinguino diversi.

Il dataset di partenza è contenuto nel file «penguins.csv».

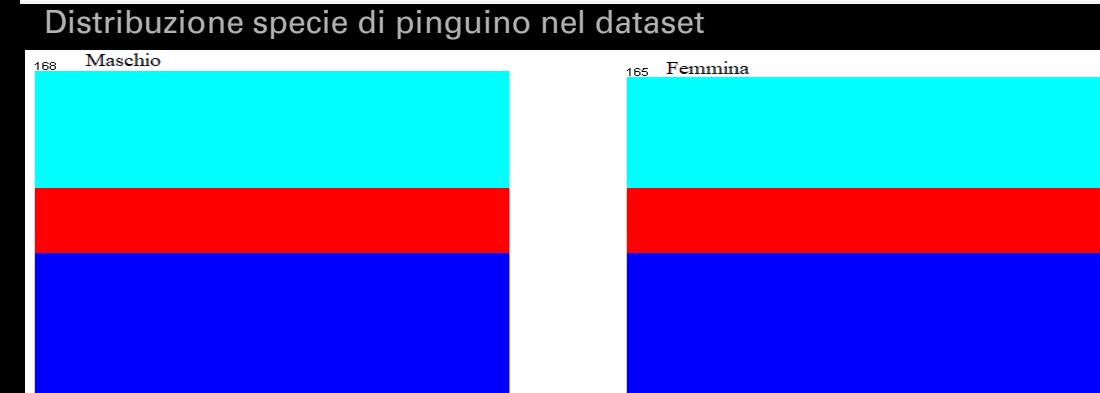
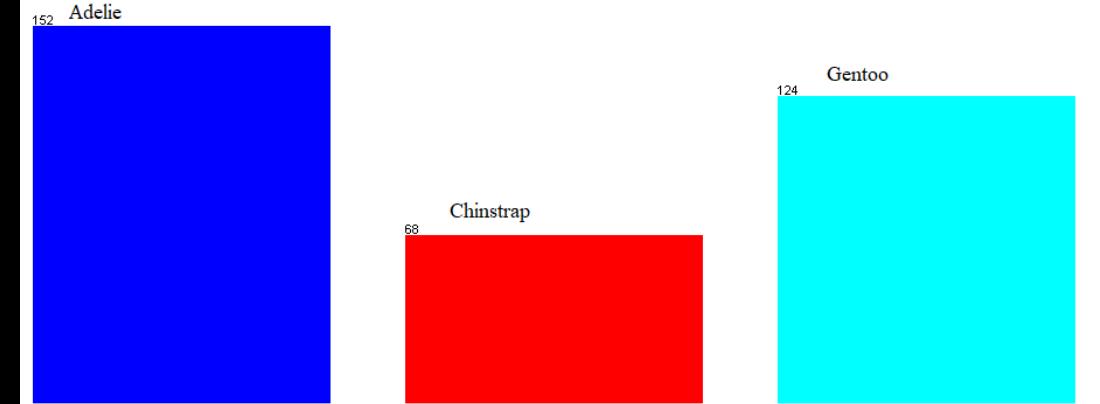
Parts of a Penguin



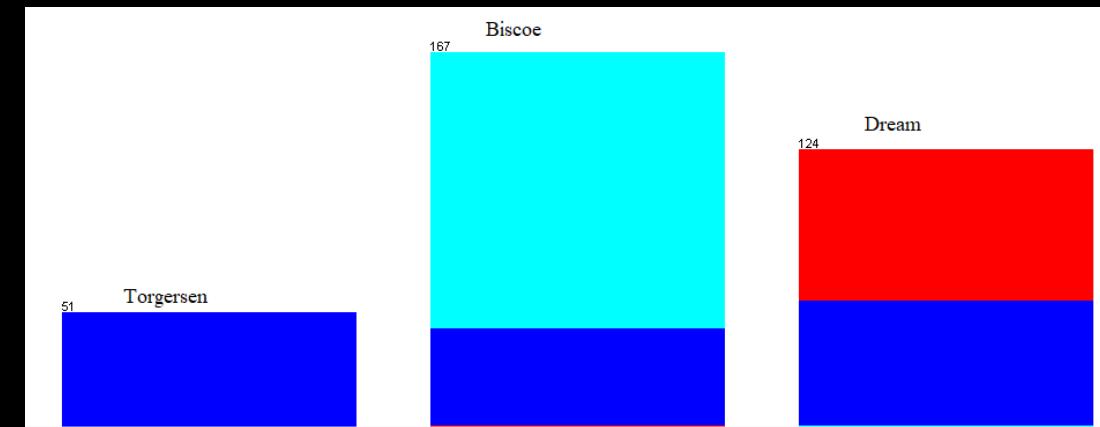
Composizione del dataset

Le specie, nel dataset in esame, non hanno la stessa frequenza, come possiamo notare nel diagramma a barre riportato di fianco. La maggior parte dei pinguini appartengono alla specie Adelie (circa il 44%), seguita dalla specie Gentoo (36%) ed infine Chinstrap risulta essere la meno frequente nel nostro dataset (circa il 20% degli esemplari). Il genere risulta invece essere ben distribuito con gli esemplari maschi che superano di tre il numero di esemplari di genere femminile come possiamo notare nel grafico riportato di fianco. Inoltre anche rispetto alla specie di appartenenza la proporzione di maschi e femmine risulta essere omogenea.

Infine, nell'ultimo diagramma a barre, possiamo notare come quasi la metà dei pinguini analizzati provenga dall'isola di Biscoe (circa il 49%), mentre da Dream e da Torgersen rispettivamente il 36% ed il 15%. Possiamo osservare inoltre che, nel dataset preso in esame, l'unica specie presente su tutte e tre le isole risulti essere Adelie mentre Gentoo è presente solo sull'isola di Biscoe ed il Chinstrap esclusivamente su Dream.



Distribuzione del genere con evidenziata la proporzione della specie di appartenenza per entrambi i generi

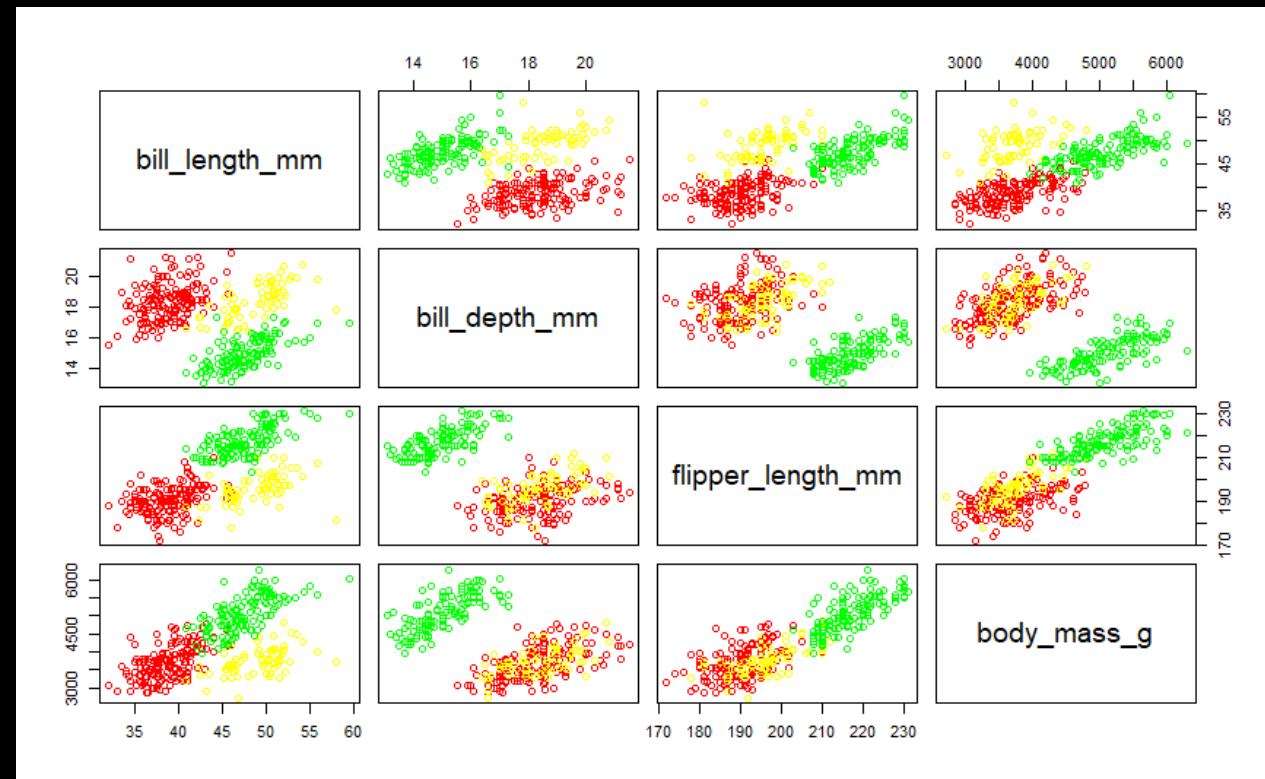


Distribuzione dell'isola di provenienza con evidenziata la proporzione della specie di appartenenza per ogni isola di provenienza

Analisi descrittiva attributi continui

Prendiamo ora in esami gli attributi continui presenti nel nostro dataset, in primo luogo rappresentiamo i nostri dati attraverso uno scatterplot riportato di fianco, ottenuto grazie al comando «plot» di R, da cui possiamo osservare come gli esemplari appartenenti alla specie Gentoo (in verde) risultino essere ben separati rispetto alle altre due specie che sembrano invece sovrapporsi. In particolare la specie «Gentoo» sembra avere una massa corporea ed una lunghezza delle ali molto maggiori rispetto alle altre due specie. Questo potrebbe quindi indurci a provare ad usare l'algoritmo di clustering Kmeans per raggruppare in cluster il nostro dataset.

Nell'immagine sotto invece sono riportate alcune caratteristiche dei quattro attributi continui, ossia: il minimo, i quartili, la media ed il massimo.



Scatterplot degli attributi continui del dataset, in rosso gli esemplari appartenenti alla specie Adelie, in giallo Chinstrap ed in verde Gentoo

bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Min. :32.10	Min. :13.10	Min. :172.0	Min. :2700
1st Qu.:39.23	1st Qu.:15.60	1st Qu.:190.0	1st Qu.:3550
Median :44.45	Median :17.30	Median :197.0	Median :4050
Mean :43.92	Mean :17.15	Mean :200.9	Mean :4202
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

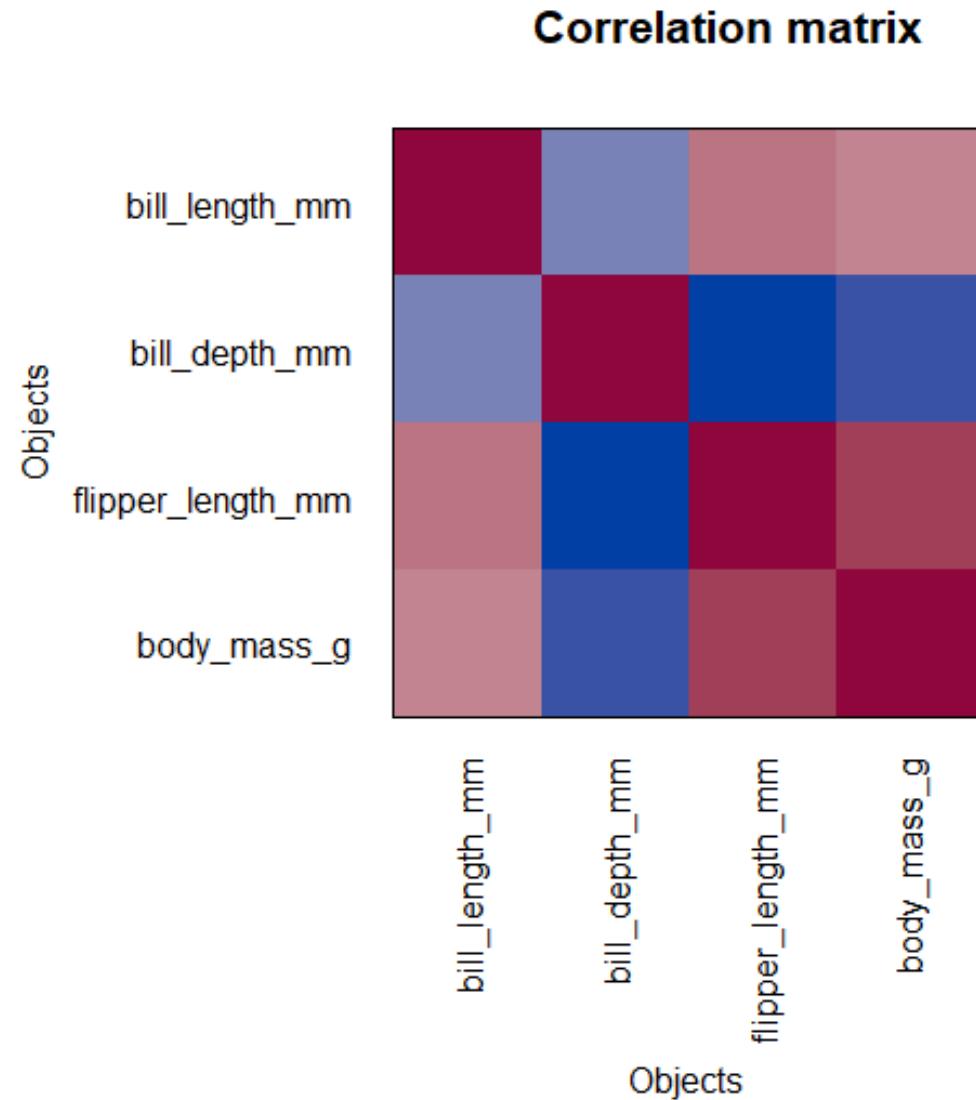
Correlazione tra gli attributi continui

Ora prendiamo invece in considerazione la correlazione tra gli attributi continui del dataset ottenuta grazie al comando «corr» di R.

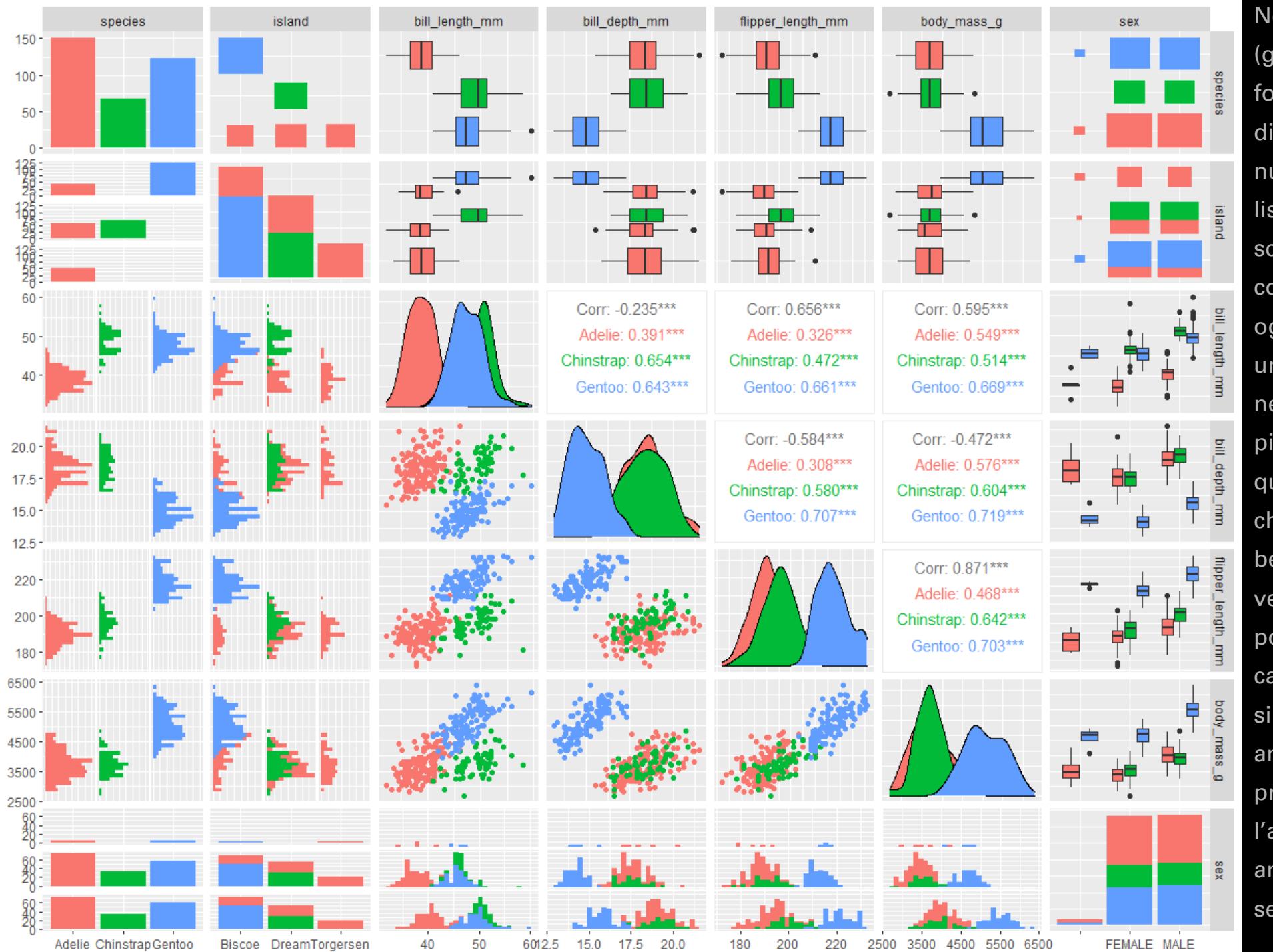
Nella tabella sotto riportata, che mostra il coefficiente di correlazione di Pearson tra le variabili, possiamo osservare come si abbia una forte correlazione positiva tra massa corporea e lunghezza delle ali (flipper), spiegabile in parte dal fatto che i pinguini più robusti necessitino di ali più lunghe per poter nuotare e quindi cacciare i pesci dei quali si nutrono.

Inoltre possiamo notare una correlazione moderatamente positiva tra lunghezza del becco e delle ali e tra lunghezza del becco e massa corporea insieme ad una correlazione moderata negativa tra lunghezza delle pinne e profondità del becco

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
bill_length_mm	1.0000000	-0.2350529	0.6561813	0.5951098
bill_depth_mm	-0.2350529	1.0000000	-0.5838512	-0.4719156
flipper_length_mm	0.6561813	-0.5838512	1.0000000	0.8712018
body_mass_g	0.5951098	-0.4719156	0.8712018	1.0000000



Matrice di correlazione che evidenzia con colori tendenti al rosso correlazioni positive mentre in blu le variabili correlate negativamente, una colorazione più scura indica una correlazione più forte



Nell'immagine mostrata di fianco (generata grazie al comando «`ggpairs`» fornito dalla libreria «`GGally`» disponibile su R) sono riportati numerosi grafici, tra i quali: istogrammi lisciati, boxplot, diagrammi a barre e scatterplot. Inoltre sono indicate la correlazione tra le variabili continue per ogni specie in esame. Possiamo notare un esempio del paradosso di Simpson nella correlazione tra lunghezza delle pinne e profondità del becco e tra quest'ultima e la massa corporea oltre che tra profondità e lunghezza del becco. Il paradosso di Simpson si verifica quando una correlazione positiva o negativa tra due variabili cambia di segno se calcolata per singoli sottogruppi del dataset in analisi. Possiamo notare anche la presenza di valori mancanti per l'attributo riguardante il genere, che andremo a trattare nella prossima sezione

Grafico delle coordinate parallele

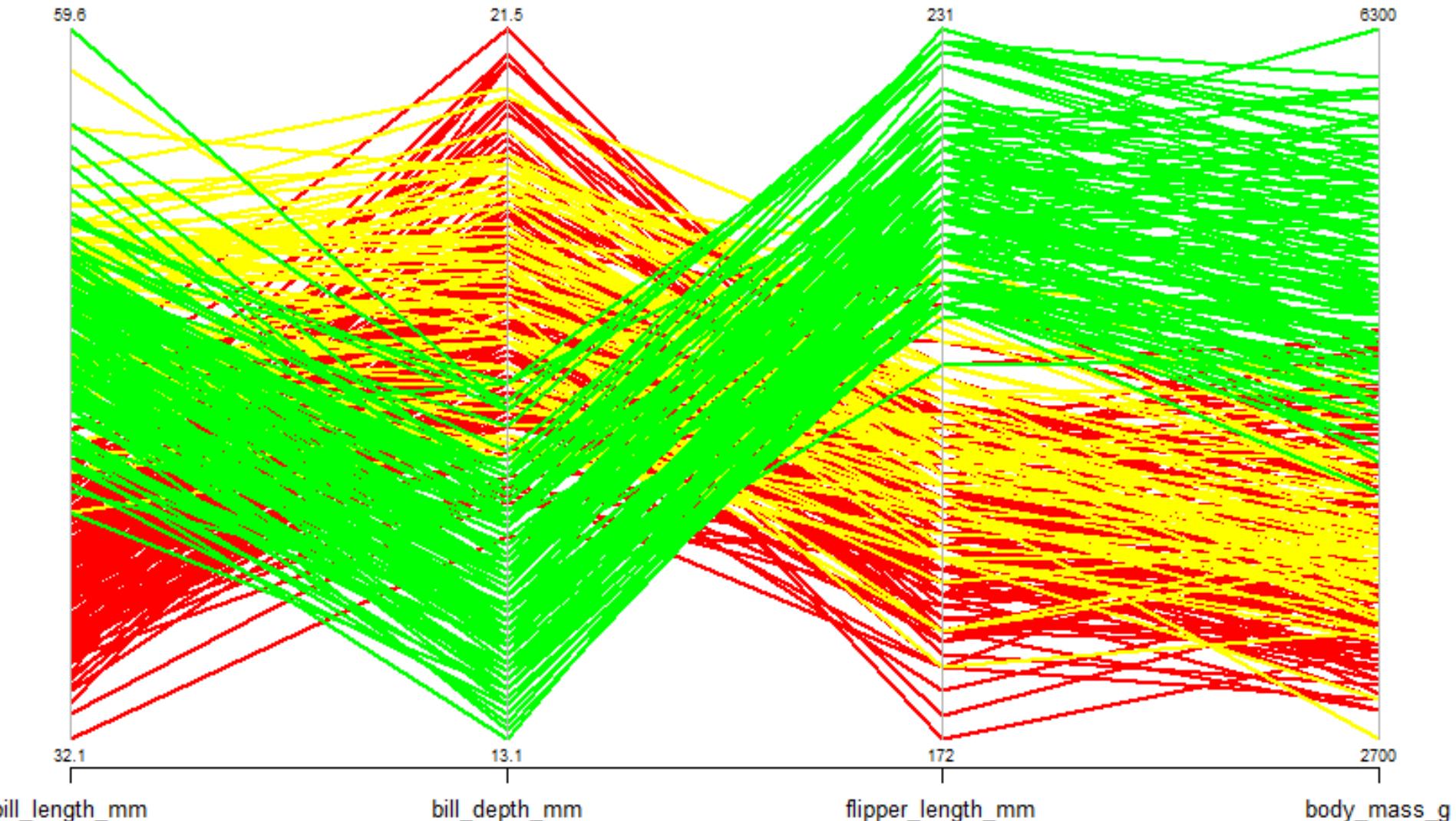
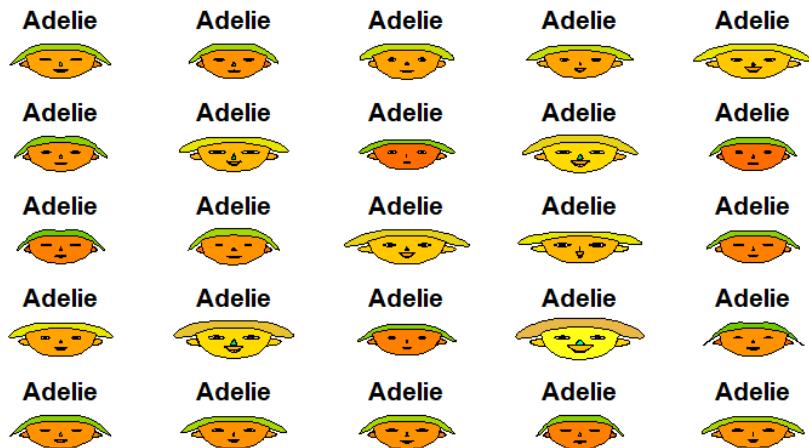


GRAFICO A COORDINATE PARALLELE DEGLI ATTRIBUTI CONTINUI DEL NOSTRO DATASET
IN VERDE SONO RIPORTATI GLI ESEMPLARI APPARTENENTI ALLA SPECIE GENTOO, IN ROSSO ADELIE MENTRE IN GIALLO CHINSTRAP. NOTIAMO COME LA SPECIE GENTOO RISULTI BEN SEPARATA RISPETTO ALLE ALTRE DUE SPECIE CHE INVECE TENDONO A CONFONDERSI E COME LA LORO MASSA CORPOREA E LUNGHEZZA DELLE PINNE RISULTI NETTAMENTE MAGGIORE CON UN BECCO NETTAMENTE meno profondo.



effect of variables:

```

modified item      Var
"height of face" "bill_length_mm"
"width of face"   "bill_depth_mm"
"structure of face" "flipper_length_mm"
"height of mouth" "body_mass_g"
"width of mouth"  "bill_length_mm"
"smiling"         "bill_depth_mm"
"height of eyes"  "flipper_length_mm"
"width of eyes"   "body_mass_g"
"height of hair"  "bill_length_mm"
"width of hair"   "bill_depth_mm"
"style of hair"   "flipper_length_mm"
"height of nose"  "body_mass_g"
"width of nose"   "bill_length_mm"
"width of ear"    "bill_depth_mm"
"height of ear"   "flipper_length_mm"

```



GRAFICO DELLE FACCE DI CHERNOFF PER UN SOTTOINSIEME DEL NOSTRO DATASET, CON IN BASSO A SINISTRA RIPORTATI GLI EFFETTI CHE GLI ATTRIBUTI HANNO SULLE VARIE CARATTERISTICHE DELLE FACCE. IN PARTICOLARE POSSIAMO NOTARE COME LE FACCE CHE CORRISPONDONO AD ESEMPLARI DELLA STESSA SPECIE TENDANO AD ASSOMIGLIARSI MOLTO, MENTRE SE CONFRONTIAMO ESEMPLARI DI SPECIE DIVERSE ESSI AVRANNO FACCE MOLTO DIFFERENTI, IN PARTICOLARE NOTIAMO COME I PINGUINI «GENTOO» TENDANO, ANCORA UNA VOLTA, AD ESSERE BEN DISTINTI DALLE ALTRE DUE SPECIE LE CUI FACCE INVECE TENDONO AD ESSERE LEGGERMENTE SIMILI.

Preprocessing del dataset

Si è reso necessario poiché il dataset presenta dati mancanti che vanno trattati per poter effettuare lo studio in maniera ottimale.

Caricamento su base di dati

Per poter visualizzare e trattare i dati mancanti in maniera efficiente è stato scelto di utilizzare una base di dati relazionale ed in particolare abbiamo sfruttato MySQL attraverso Workbench. In primo luogo si è reso necessario creare una tabella, chiamata «penguins» su cui importare i dati presenti nel file CSV di partenza, questo procedimento è stato effettuato grazie al wizard di import automatico fornito da Workbench. Per poter individuare in maniera più semplice i dati mancanti abbiamo definito tutti gli attributi come di tipo testuale (in questa fase non saranno infatti necessarie operazioni di tipo algebrico). Una volta completato il caricamento , per trovare i dati mancanti abbiamo usato la query:

```
SELECT * FROM datamining.penguins where sex='' or species='' or island='' or bill_length_mm='' or bill_depth_mm='' or body_mass_g='' or flipper_length_mm=''.
```

Individuando così 11 pinguini che presentavano attributi mancanti. Di questi 11 esemplari, 9 erano privi dell'attributo riguardante il genere, mentre per gli altri 2 l'unica informazione presente era la specie e l'isola di provenienza. È stato quindi deciso di rimuovere dal dataset questi due esemplari in quanto non forniscono quasi nessuna informazione aggiuntiva rispetto al resto dei pinguini. Il resto degli esemplari è stato invece salvato in un nuovo file CSV che verrà caricato su Weka per la fase successiva di preprocessing.

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
	Adelie	Torgersen					
	Adelie	Torgersen	34.1	18.1	193	3475	
	Adelie	Torgersen	42	20.2	190	4250	
	Adelie	Torgersen	37.8	17.1	186	3300	
	Adelie	Torgersen	37.8	17.3	180	3700	
	Adelie	Dream	37.5	18.9	179	2975	
	Gentoo	Biscoe	44.5	14.3	216	4100	
	Gentoo	Biscoe	46.2	14.4	214	4650	
	Gentoo	Biscoe	47.3	13.8	216	4725	
	Gentoo	Biscoe	44.5	15.7	217	4875	
▶	Gentoo	Biscoe					

ESEMPLARI CHE PRESENTANO DATI MANCANTI VISUALIZZATI IN
MYSQL WORKBENCH

Preprocessing su Weka

Abbiamo quindi caricato il dataset parzialmente pulito sul software Weka per poter ultimare il lavoro di pulizia dei dati. Mancano infatti da trattare gli ultimi nove esemplari dei quali non è stato registrato il genere. Per fare ciò abbiamo quindi sfruttato il filtro «ReplaceMissingValue», disponibile su Weka, che ha sostituito i valori mancanti con la moda dell'attributo categorico mancante, nel nostro caso l'attributo genere che è stato posto a maschio per tutti i pinguini con l'attributo genere non specificato. Abbiamo quindi rimpiazzato tutti i valori mancanti e siamo quindi pronti per effettuare il nostro studio avendo rimpiazzato tutti i valori mancanti nel nostro dataset.

Il dataset così ottenuto è stato salvato nel file CSV «penguins_puliti.csv».

Clustering

La prima analisi che abbiamo deciso di effettuare è quella dei cluster sfruttando in particolare gli algoritmi di clustering K-means e quello gerarchico agglomerativo entrambi implementati in Weka

Clustering K-means

Nelle sezioni precedenti, analizzando gli scatterplot degli attributi continui, avevamo osservato come le tre specie di pinguino sembrassero essere abbastanza distinte tra loro, ed in particolare la specie Gentoo sembrava essere particolarmente «separata» dalle altre due specie che invece sembravano essere più sovrapposte tra loro.

Per verificare queste osservazioni è stato deciso in primo luogo di effettuare un analisi dei cluster sfruttando l'algoritmo di clustering K-means, che nel nostro caso dovrebbe essere abbastanza soddisfacente in quanto in primo luogo conosciamo il numero di cluster che desideriamo ottenere dall'esecuzione dell'algoritmo, ma anche perché i tre gruppi individuati dalle specie degli esemplari risultano ad occhio essere abbastanza separati tra loro oltre a possedere una forma non troppo «bizzarra».

Non dobbiamo però dimenticare che l'esecuzione di tale algoritmo è non deterministica in quanto varia in base alla scelta dei centroidi di partenza, nel nostro caso quindi abbiamo testato vari «seed» per la generazione dei centroidi iniziali e abbiamo preso in considerazione quello che ha fornito un risultato «migliore».

Risultati dell'applicazione dell'algoritmo K-means

In primo luogo abbiamo testato l'algoritmo di clustering K-means, implementato in Weka con il nome «SimpleKmeans», sui dati non normalizzati ma sfruttando la distanza euclidea con l'opzione «Don't normalize» posta a falso, ignorando gli attributi categorici del dataset, e ponendo il numero di cluster pari a 3.

Impostando come seme di generazione 10 il risultato ottenuto risulta essere soddisfacente con appena 11% dei dati raggruppati in maniera incorretta.

Possiamo inoltre notare, come avevamo previsto in precedenza, che nessun esemplare appartenente alla specie Gentoo è stato inserito in un gruppo diverso dalla specie di appartenenza, inoltre viceversa nessun pinguino Chinstrap o Adelie è stato incorrettamente inserito nel gruppo dei pinguini Gentoo. Ciò dimostra che la specie Gentoo risulta effettivamente ben separata rispetto alle altre due specie che invece hanno alcuni esemplari inseriti nei cluster che rappresentano la specie errata, in particolare ben 31 esemplari della specie Adelie sono stati inseriti nel cluster contenente pinguini della specie Chinstrap. Ciò mostra quindi che, almeno alcuni esemplari di queste due specie, non risultano avere i valori attribuiti in esame troppo distanti tra loro.

Final cluster centroids:					
Attribute	Full Data (342.0)	Cluster#			
		0 (91.0)	1 (128.0)	2 (123.0)	
bill_length_mm	43.9219	47.0198	38.2766	47.5049	
bill_depth_mm	17.1512	18.8769	18.0086	14.9821	
flipper_length_mm	200.9152	197.1868	187.9297	217.187	
body_mass_g	4201.7544	3948.0769	3541.9922	5076.0163	

CENTROIDI FINALI DEI TRE CLUSTER RISULTANTI

```
==== Model and evaluation on training set ===

Clustered Instances

0      91 ( 27%)
1     128 ( 37%)
2     123 ( 36%)

Class attribute: species
Classes to Clusters:

0   1   2  <-- assigned to cluster
31 120  0 | Adelie
60   8  0 | Chinstrap
0   0 123 | Gentoo

Cluster 0 <-- Chinstrap
Cluster 1 <-- Adelie
Cluster 2 <-- Gentoo

Incorrectly clustered instances :      39.0      11.4035 %
```

Risultati dell'applicazione dell'algoritmo K-means su dati normalizzati

Successivamente abbiamo provato a sfruttare la normalizzazione Z-score, offerta dal filtro «Standardize» di Weka, degli attributi continui del nostro dataset per osservare se ciò riusciva a migliorare le prestazioni dell'algoritmo di clustering K-means visto prima.

Sono stati mantenuti gli stessi parametri con cui avevamo eseguito l'algoritmo di clustering Kmeans in precedenza, ponendo però in questo caso l'opzione «Don't normalize» a vera per il calcolo della distanza euclidea.

Abbiamo quindi verificato che in questo caso l'algoritmo di clustering è riuscito a raggruppare in maniera più precisa gli esemplari nei tre cluster rappresentanti le tre specie. Infatti la percentuale di istanze raggruppate in maniera incorretta risulta essere appena del 8,5% circa, con ben 10 pinguini raggruppati correttamente in più rispetto all'esecuzione precedente.

Nella diapositiva seguente sono riportati quindi gli scatterplot che evidenziano i gruppi ricavati da quest'esecuzione dell'algoritmo di clustering con gli esemplari inseriti in cluster incorretti evidenziati dal fatto di avere l'icona a forma di quadrato.

Final cluster centroids:				
Attribute	Full Data (342.0)	Cluster#		
		0 (87.0)	1 (132.0)	2 (123.0)
<hr/>				
bill_length_mm	0	0.66	-1.0465	0.6563
bill_depth_mm	0	0.8157	0.4858	-1.0984
flipper_length_mm	-0	-0.2858	-0.8899	1.1572
body_mass_g	-0	-0.3738	-0.7695	1.0902

```
== Model and evaluation on training set ==

Clustered Instances

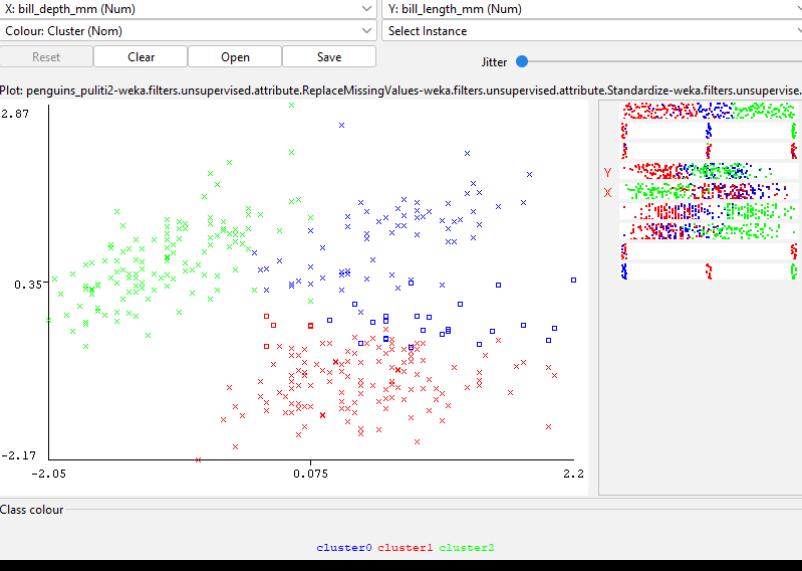
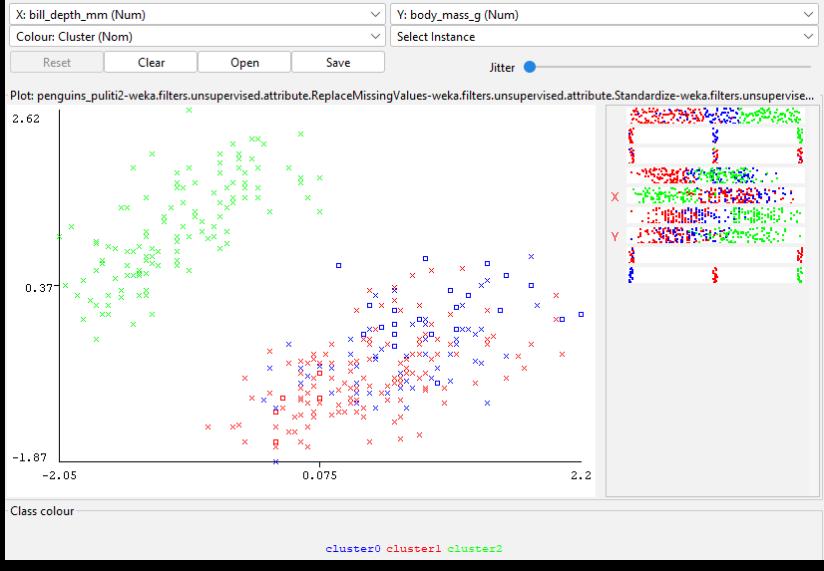
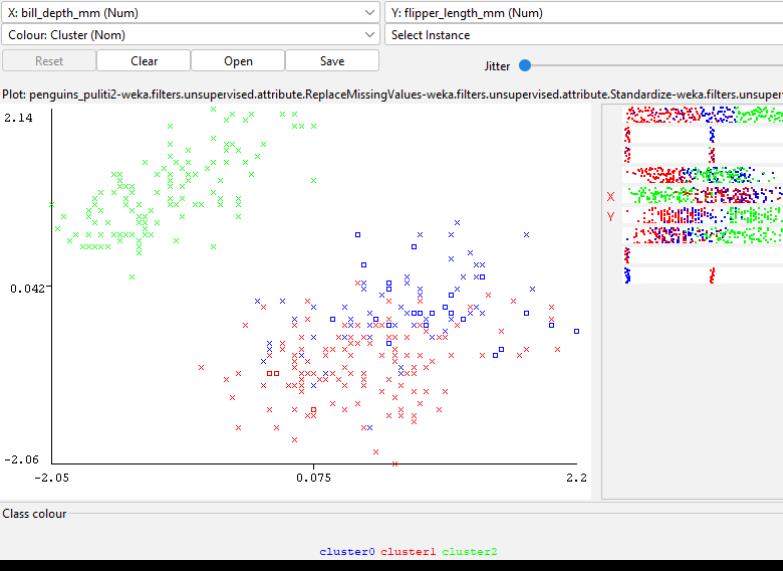
0      87 ( 25%)
1      132 ( 39%)
2      123 ( 36%)

Class attribute: species
Classes to Clusters:

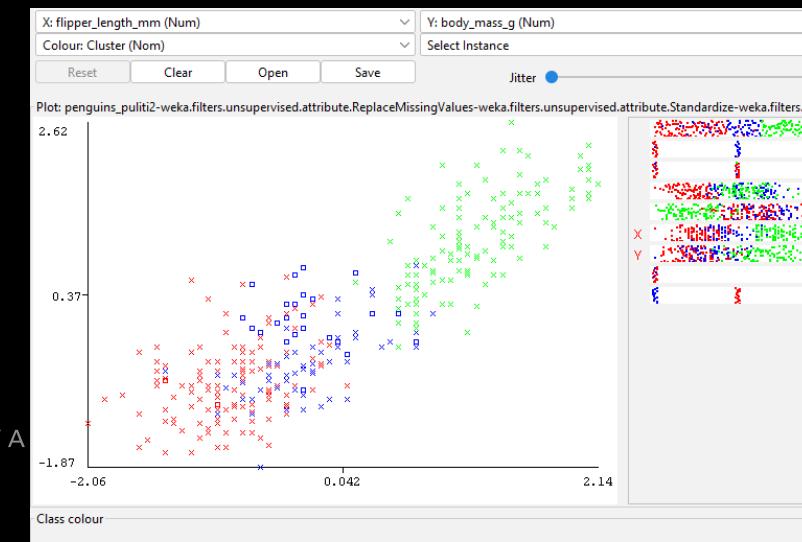
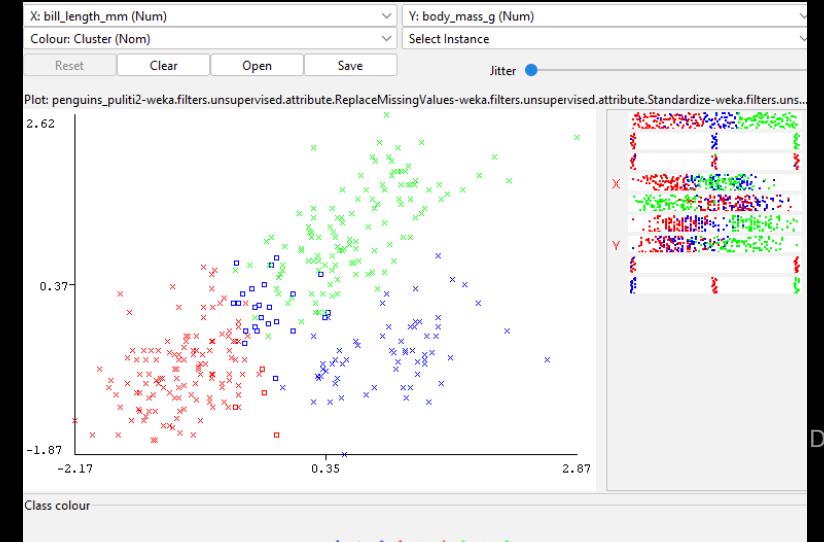
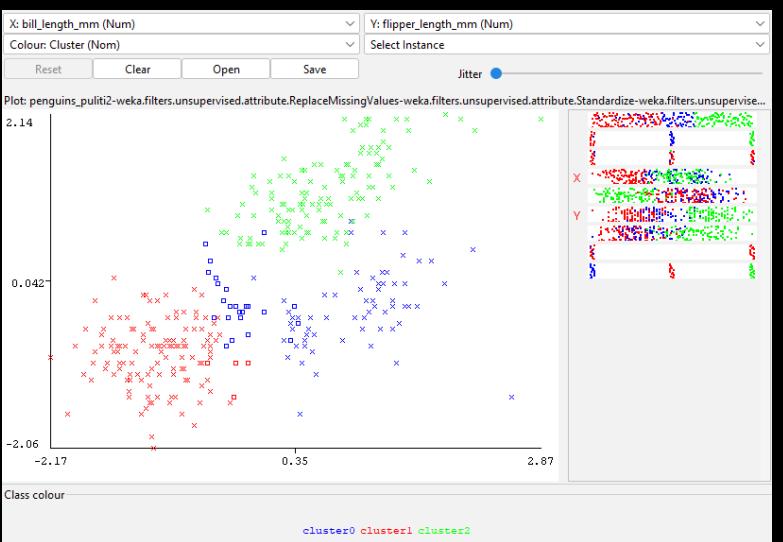
0 1 2 <-- assigned to cluster
24 127 0 | Adelie
63 5 0 | Chinstrap
0 0 123 | Gentoo

Cluster 0 <-- Chinstrap
Cluster 1 <-- Adelie
Cluster 2 <-- Gentoo

Incorrectly clustered instances : 29.0    8.4795 %
```



Notiamo come i pinguini raggruppati nel cluster che rappresenta la specie Gentoo (cluster 2 in verde) siano effettivamente ben separati ed il cluster 2 risulti quindi privo di pinguini di specie diverse, mentre gli altri due cluster rappresentanti le altre due specie, Chinstrap (cluster 0 in blu) e Adelie (cluster 1 in rosso) contengono anche esemplari appartenenti all'altra specie.



DELL'A

Clustering gerarchico agglomerativo mean link

Oltre al clustering effettuato tramite l'algoritmo K-means, abbiamo deciso di sfruttare anche l'algoritmo di clustering gerarchico agglomerativo usando il metodo del legame medio per definire la distanza tra due cluster. Per fare ciò è stato sfruttato il metodo «HierarchicalClusterer» di Weka con l'opzione linkType impostata a MEAN e numero di cluster pari a 3.

L'algoritmo di clustering gerarchico agglomerativo è stato eseguito con tutti i metodi di calcolo della distanza tra cluster disponibili in Weka, ma il metodo del legame medio è risultato essere il migliore tra tutti ed è stato quindi scelto per la nostra analisi.

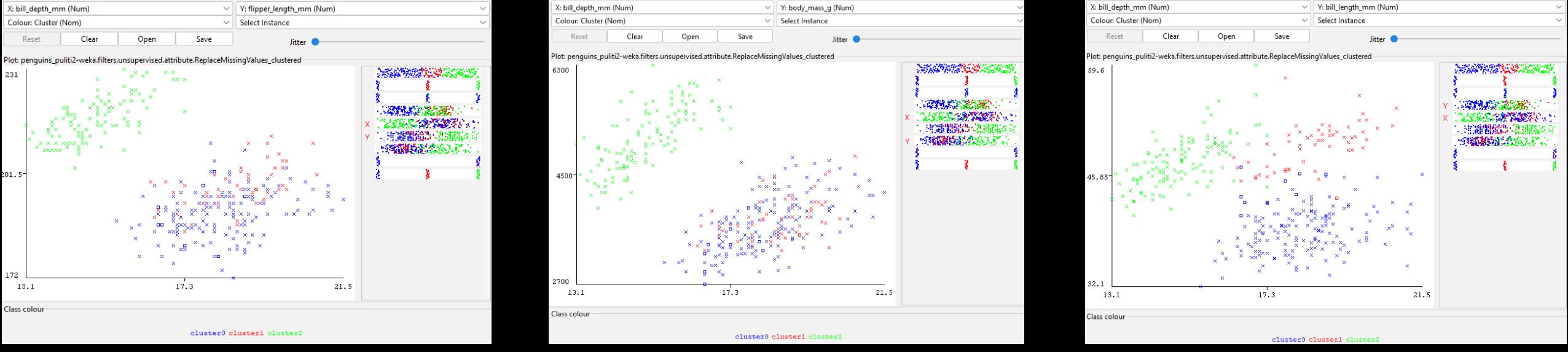
Come possiamo notare dall'immagine riportata a fianco, il risultato dell'esecuzione è ottimo in quanto appena il 3% degli esemplari è stato inserito in un gruppo corrispondente alla specie sbagliata. Inoltre osserviamo, come già avvenuto con l'algoritmo di clustering K-means, che il cluster 2, contenente tutti i pinguini appartenenti alla specie Gentoo, non contiene nessun'altro pinguino di specie diversa. Questo dimostra ulteriormente come la specie Gentoo, almeno per quanto riguarda gli attributi osservati, risulta essere notevolmente diversa e «separata» rispetto alle altre due specie di pinguino. Nelle pagine seguenti riportiamo quindi sia gli scatterplot che mostrano il cluster a cui ogni esemplare è stato assegnato ed il dendogramma per visualizzare il processo di agglomerazione delle istanze tagliato in modo da avere tre gruppi risultanti dal esecuzione dell'algoritmo

```
Clustered Instances
0      160 ( 47%)
1       59 ( 17%)
2      123 ( 36%)

Class attribute: species
Classes to Clusters:
0   1   2 <-- assigned to cluster
150   1   0 | Adelie
 10  58   0 | Chinstrap
   0   0 123 | Gentoo

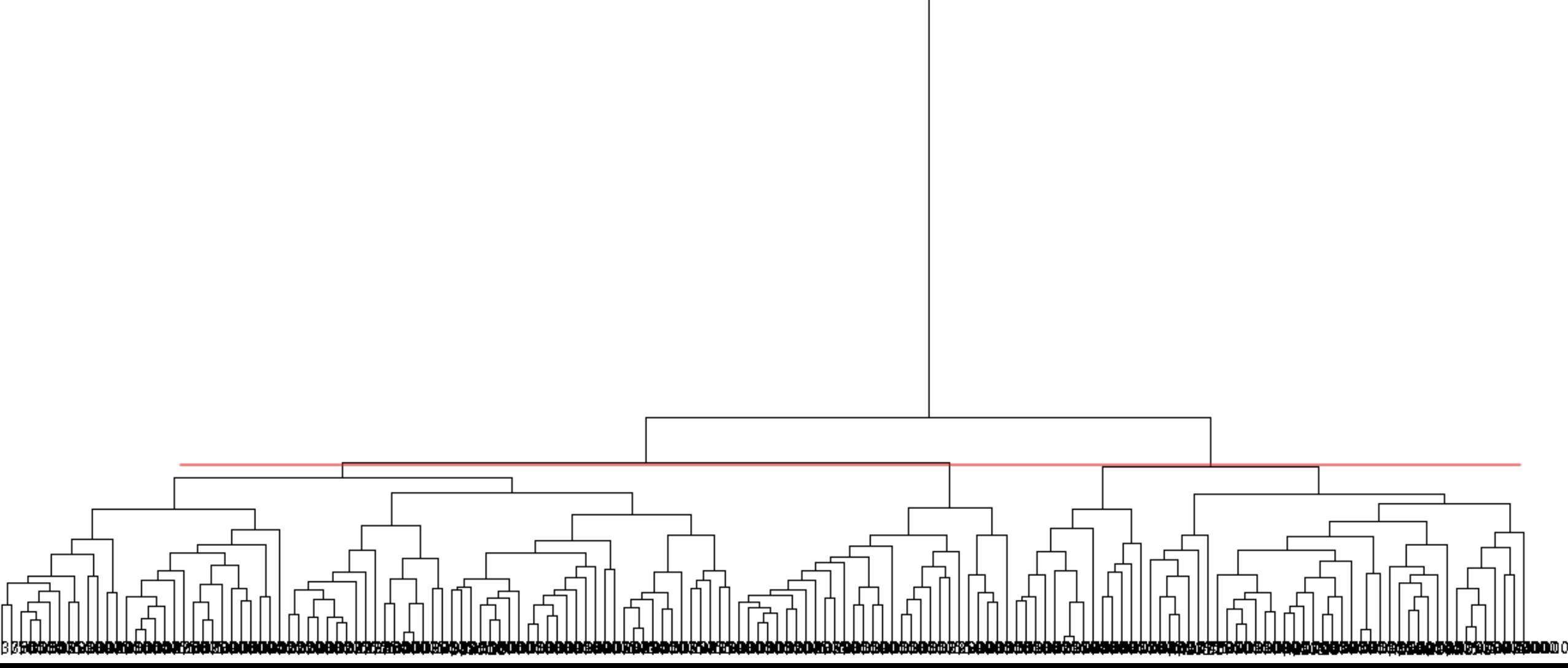
Cluster 0 <-- Adelie
Cluster 1 <-- Chinstrap
Cluster 2 <-- Gentoo

Incorrectly clustered instances :           11.0      3.2164 %
```



In questo caso il cluster 0 (colore blu) raggruppa i pinguini della specie Adelie, il cluster 1 (rosso) quelli appartenenti alla specie Chinstrap mentre il cluster 2 (verde) contiene pinguini della specie Gentoo.
 Notiamo come gli esemplari raggruppati in maniera incorretta (rappresentati col simbolo di un quadrato) siano in numero inferiore rispetto a quelli presenti negli scatterplot risultanti dall'applicazione dell'algoritmo di clustering K-means visti in precedenza.





Dendogramma tagliato rappresentante le varie fusioni dei cluster effettuate durante l'esecuzione dell'algoritmo col metodo del legame medio. Purtroppo disponendo di così tanti esemplari, il dendogramma risulta di difficile comprensione, anche se possiamo notare come il gruppo a destra risultante dal taglio del dendogramma (contenente esclusivamente pinguini della specie Gentoo) risulti avere una distanza di fusione molto ampia con il cluster risultante dalla fusione degli altri due cluster rimasti, che invece hanno una distanza di fusione nettamente inferiore, ciò conferma ulteriormente la tesi che i pinguini appartenenti alla specie Gentoo siano nettamente distinti rispetto alle altre due specie.

Commento finale

Abbiamo quindi avuto conferma di quello che avevamo visualizzato in precedenza nella parte di Data Understanding, ossia che la specie di pinguini «Gentoo» si differenzia in maniera sostanziale rispetto alle altre due specie studiate, almeno per quanto riguarda gli attributi in esame, poiché, con tutti gli algoritmi di clustering che abbiamo utilizzato, tale gruppo è stato sempre correttamente separato dalle altre due specie, che invece tendono ad essere sovrapposte anche se in minima parte.

Abbiamo inoltre trovato quindi che il miglior algoritmo di clustering per il nostro dataset è quello gerarchico agglomerativo con il metodo del legame medio («mean» in Weka) per il calcolo della distanza tra cluster, anche se il clustering K-means resta comunque valido come algoritmo di clustering per il nostro dataset visto che riesce comunque a «catturare» i cluster naturali rappresentati dalle specie in maniera abbastanza soddisfacente, in particolare con dati normalizzati.

Classificazione

Vogliamo trovare il modello di classificazione migliore che ci permetta di classificare un esemplare di pinguino dati gli attributi studiati in precedenza. Abbiamo usato come attributo classe prima la specie di appartenenza e successivamente il genere.

Classificazione basata sulle specie

Partiamo quindi dal classificare gli esemplari in base alla specie di appartenenza

Classificazione delle specie tramite alberi di decisione

Come prima tecnica di classificazione è stata scelta quella che sfrutta gli alberi di decisione, usando in particolare il metodo J48 implementato in Weka. Per il primo tentativo sono stati mantenuti i parametri di default ed è stato eseguito come metodo di valutazione della performance una cross validation che divide il dataset in 10 parti.

Come possiamo notare nella prossima slide, l'albero di decisione derivante risulta però essere particolarmente ampio e profondo, rendendolo di difficile comprensione e interpretazione, nonostante la accuratezza risultati molto soddisfacente (96,2%).

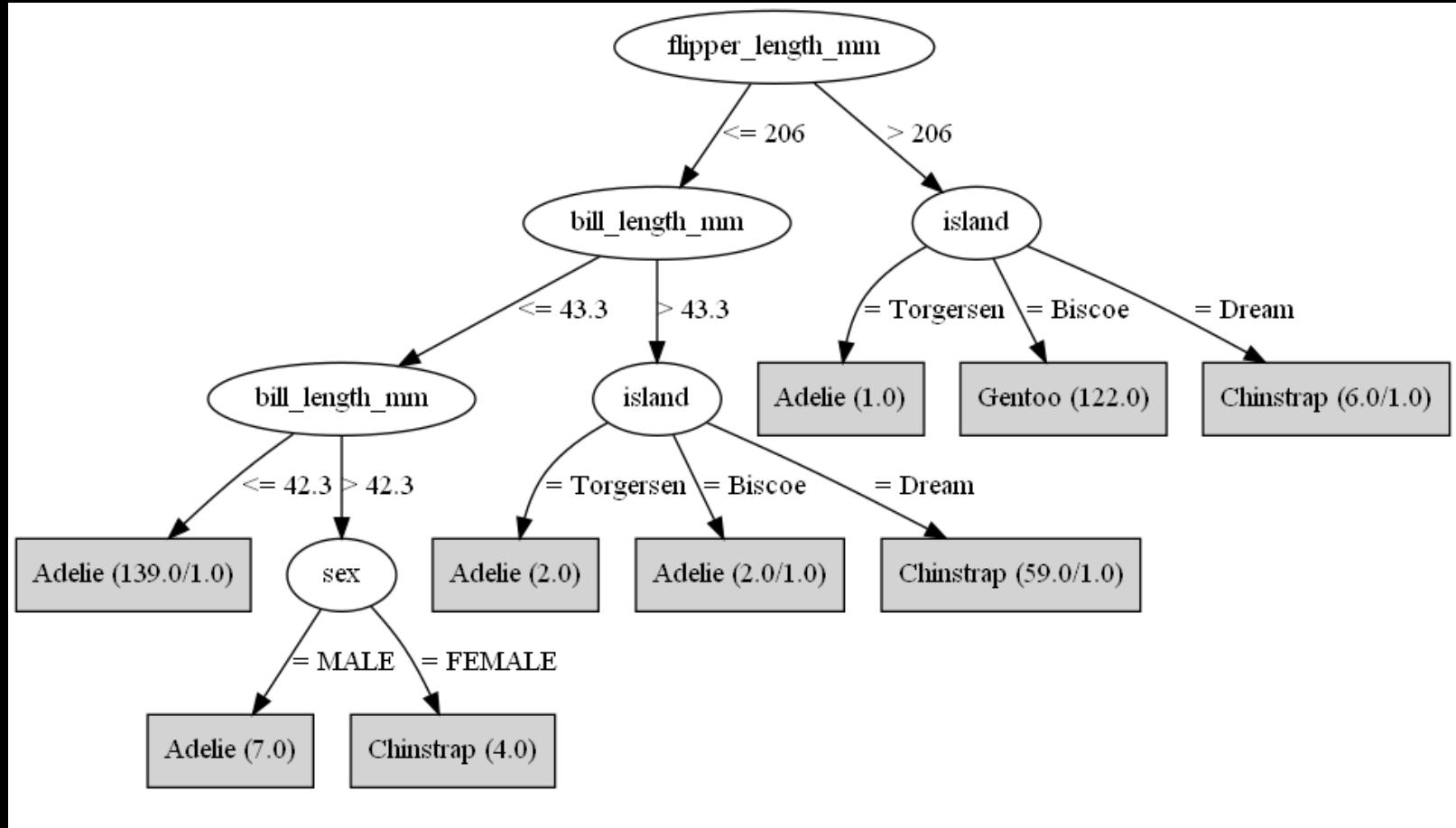
Per questo è stato deciso di aumentare il numero minimo di esemplari per foglia dell'albero da 2 (valore di default) a 10, sacrificando un minimo di accuratezza per ottenere un albero più «semplice» e comprensibile.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,954	0,026	0,966	0,954	0,960	0,929	0,967	0,955	Adelie
	0,956	0,015	0,942	0,956	0,949	0,936	0,970	0,890	Chinstrap
	0,976	0,018	0,968	0,976	0,972	0,956	0,992	0,982	Gentoo
Weighted Avg.	0,962	0,021	0,962	0,962	0,962	0,940	0,976	0,952	

==== Confusion Matrix ===

a	b	c	<-- classified as
144	4	3	a = Adelie
2	65	1	b = Chinstrap
3	0	120	c = Gentoo

*Notiamo come
l'albero abbia un
altezza di 4, con ben
9 foglie separate e
sei nodi interni
rappresentanti
ciascuno un
confronto diverso,
per questo è stato
deciso di aumentare
il numero di
esemplari per foglia
a 10*



Classificazione delle specie tramite alberi di decisione

01/2023

STUDIO DELLE SPECIE DI PINGUINI PRESENTI
DELL'ARCTICA E PALMERA

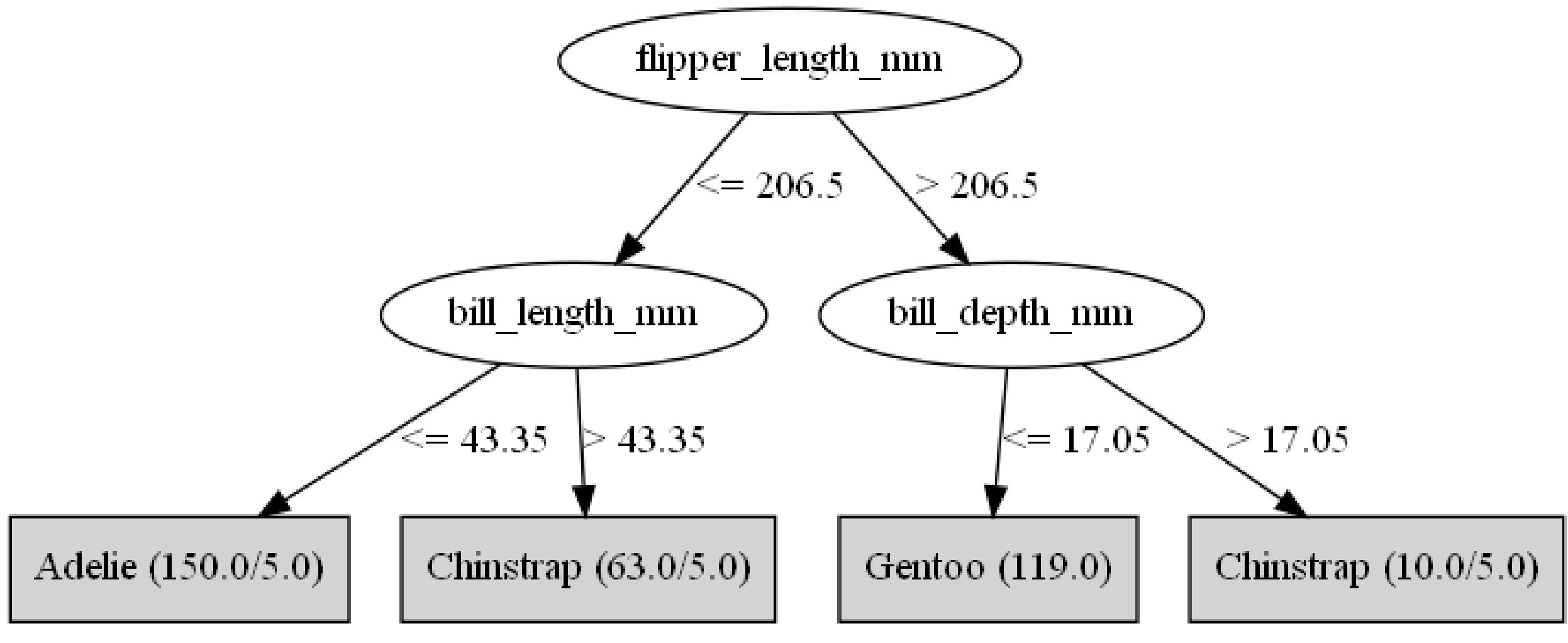
È stato quindi rieseguito il metodo J48 impostando a 10 il numero minimo di esemplari per foglia e ponendo a vero l'opzione per fare in modo che si possa avere come valore di split di un attributo continuo anche un valore non presente nel dataset. Per quanto riguarda invece la valutazione della performance è stata utilizzata sempre la cross validation con divisione del dataset in 10 parti.

L'albero risultante, riportato nel lucido successivo, risulta essere molto più semplice rispetto a quello visto in precedenza, avendo però sacrificato circa un 3% di accuratezza, che rimane comunque del 92,6% che è un valore nonostante tutto molto elevato.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,940	0,058	0,928	0,940	0,934	0,882	0,969	0,950	Adelie
	0,794	0,026	0,885	0,794	0,837	0,801	0,923	0,853	Chinstrap
	0,984	0,032	0,945	0,984	0,964	0,944	0,981	0,960	Gentoo
Weighted Avg.	0,927	0,042	0,926	0,927	0,926	0,888	0,964	0,935	

==== Confusion Matrix ===

a	b	c	<-- classified as
142	7	2	a = Adelie
9	54	5	b = Chinstrap
2	0	121	c = Gentoo



Notiamo come l'albero abbia un'altezza pari a 2, con 4 foglie e 3 nodi interni compresa la radice rappresentanti ciascuno un confronto diverso da effettuare sull'esemplare da classificare, per classificare un esemplare eseguirò quindi sempre due confronti, il primo sull'attributo riguardante la lunghezza delle ali ed il secondo o la lunghezza o la profondità del becco.

Osserviamo inoltre come non sia stato usato alcun attributo di tipo categorico come l'isola di provenienza o il genere, sfruttando solo gli attributi continui del nostro dataset per classificare gli esemplari.

Classificazione delle specie basata su regole

01/2023

Successivamente abbiamo deciso di sfruttare la tecnica di classificazione basata su regole, in particolare tramite il metodo «JRip» implementato in Weka. Sono stati lasciati tutti i valori di default ed è stata utilizzato come metodo di valutazione della performance al solito la cross validation in 10 parti. Le regole risultanti sono riportate di seguito:

```
(island = Dream) and (bill_length_mm >= 42.4) => species=Chinstrap (69.0/2.0)
(flipper_length_mm >= 207) and (island = Biscoe) => species=Gentoo (122.0/0.0)
=> species=Adelie (151.0/2.0)
```

Notiamo come siano state generate appena due regole più una di default, ossia il numero minimo possibile visto che il mio attributo classe (la specie) ha cardinalità pari a tre. Ciascuna regola contiene, a differenza di quanto avvenuto in precedenza con gli alberi di decisione, anche una condizione basata sull'attributo dell'isola di provenienza. Grazie anche a questo l'accuratezza risultante è addirittura del 97,7%, superando entrambi gli alberi visti in precedenza con appena 8 esemplari classificati in maniera errata.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,974	0,021	0,974	0,974	0,974	0,953	0,970	0,948	Adelie
	0,971	0,007	0,971	0,971	0,971	0,963	0,982	0,927	Chinstrap
	0,984	0,009	0,984	0,984	0,984	0,975	0,987	0,973	Gentoo
Weighted Avg.	0,977	0,014	0,977	0,977	0,977	0,963	0,979	0,953	

== Confusion Matrix ==

```
a    b    c    <- classified as
147   2    2 |   a = Adelie
      2   66   0 |   b = Chinstrap
      2    0 121 |   c = Gentoo
```

STUDIO DELLE SPECIE DI PINGUINI PRESENTI
DELL'ARCIPERLA GOLFO DI PALMERA

Classificazione delle specie basata sui vicini più prossimi (Nearest Neighbours)

Come prossima tecnica di classificazione è stata selezionata quella basata sull'algoritmo dei vicini più prossimi (Nearest Neighbours), implementata in Weka nel metodo «IBk». Il numero di vicini più prossimi da prendere in considerazione è stato posto a 10 esemplari e come algoritmo di ricerca è stato mantenuto quello di default ossia quello lineare con distanza euclidea normalizzata.

I risultati, riportati in fondo, sono estremamente positivi, se consideriamo che abbiamo utilizzato come metodo di valutazione della performance la cross validation in 10 parti. Infatti abbiamo appena un esemplare classificato in maniera incorretta, con una percentuale di accuratezza del 99,7%, che è dovuta in gran parte alla separazione tra le diverse specie che avevamo osservato in precedenza nella parte dedicata al clustering, che fa in modo che i vicini più prossimi di ciascun esemplare risultino essere sempre della stessa specie.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1,000	0,005	0,993	1,000	0,997	0,994	1,000	1,000	Adelie
	0,985	0,000	1,000	0,985	0,993	0,991	1,000	0,999	Chinstr
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Gentoo
Weighted Avg.	0,997	0,002	0,997	0,997	0,997	0,996	1,000	1,000	

==== Confusion Matrix ===

a	b	c	<-- classified as
151	0	0	a = Adelie
1	67	0	b = Chinstrap
0	0	123	c = Gentoo

Classificazione delle specie basata sul teorema di Bayes

01/2023

STUDIO DELLE SPECIE DI PINGUINI PRESENTI
DELL'ARCTICA AL PALMERO

Ora sfruttiamo invece la tecnica di classificazione che si basa sull'algoritmo di Bayes, usando: lo stimatore di Laplace per stimare la probabilità condizionata associata ad un possibile valore di un attributo categorico data una specie, in modo da non poter avere valori nulli e la distribuzione normale per stimare la probabilità condizionata di un attributo continuo. Per fare ciò è stato usato il metodo «NaiveBayes» disponibile in Weka, senza cambiare i parametri di default.

I risultati mostrano un'accuratezza del 97,7%, utilizzando al solito la cross validation che divide il dataset in 10 parti, la stessa che avevamo ottenuto con la tecnica basata sulle regole, anche se i soggetti classificati in maniera incorrecta non sono esattamente gli stessi (come è possibile notare anche dalle due tabelle di confusione).

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,980	0,026	0,967	0,980	0,974	0,953	0,998	0,997	Adelie
	0,926	0,011	0,955	0,926	0,940	0,926	0,996	0,988	Chinstr
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Gentoo
Weighted Avg.	0,977	0,014	0,977	0,977	0,977	0,964	0,998	0,996	

==== Confusion Matrix ===

a	b	c	<-- classified as
148	3	0	a = Adelie
5	63	0	b = Chinstrap
0	0	123	c = Gentoo

Classificazione delle specie basata su reti neurali artificiali (ANN)

Come ultima tecnica di classificazione è stato scelto di usare quella basata su reti neurali artificiali, sfruttando come modello il percepitrone (perceptron) multilivello, implementato in Weka nel metodo «MultilayerPerceptron», che abbiamo eseguito lasciando invariati i parametri di default e testato sempre attraverso la cross validation in 10 parti.

Il risultato ottenuto dalla cross validation , anche in questo caso, risulta essere estremamente positivo, come possiamo osservare dalla tabella riportata in basso. Infatti appena un esemplare risulta classificato in maniera incorretta, con un accuratezza pari quindi al 99,7%, come era già avvenuto per la tecnica basata sui vicini più prossimi, anche se tale esemplare classificato in maniera incorretta non è lo stesso.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,993	0,000	1,000	0,993	0,997	0,994	1,000	1,000	Adelie
	1,000	0,004	0,986	1,000	0,993	0,991	1,000	1,000	Chinstrap
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Gentoo
Weighted Avg.	0,997	0,001	0,997	0,997	0,997	0,996	1,000	1,000	

==== Confusion Matrix ===

a	b	c	<-- classified as
150	1	0	a = Adelie
0	68	0	b = Chinstrap
0	0	123	c = Gentoo

Commenti classificazione delle specie

Abbiamo quindi osservato come le migliori tecniche di classificazione, ossia quelle che permettono di ottenere un'accuratezza maggiore, prendendo come classe in analisi la specie di appartenenza degli esemplari, siano risultate essere quella dei vicini più prossimi (Nearest Neighbours), con un numero di vicini presi in considerazioni pari a 10, insieme a quella basata su reti neurali artificiali, in particolare sfruttando il percepitrone multilivello. Tali tecniche infatti classificano in maniera errata appena un esemplare usando la cross validation, con un'accuratezza pari al 99,7%.

Tali tecniche hanno però lo svantaggio di non fornire un modello attraverso il quale classificare in maniera semplice gli esemplari, per questa ragione, se il nostro obiettivo è quello di ricavare un modello sufficientemente semplice da comprendere e da utilizzare, la scelta migliore potrebbe essere quella di ricorrere alla classificazione basata su regole, in quanto con appena due regole (più una di default), entrambe contenenti due condizioni da rispettare, riesce ad ottenere un'accuratezza, usando la cross validation, pari al 97,7%.

Classificazione basata sul genere

Classifichiamo ora invece gli esemplari in base al genere di appartenenza.

Preprocessing

Prima di effettuare l'analisi è stato deciso di rimuovere dal nostro dataset gli esemplari che non avevano originariamente indicato l'attributo riguardante il genere, non rischiando così di avere valori incorretti nel mio attributo classe, in modo da poter ottenere una classificazione che non risenta di eventuali errori commessi nel rimpiazzo degli attributi mancanti.

Per fare ciò abbiamo sfruttato nuovamente MySQL con la tabella «penguins», contenente tutti gli esemplari, che avevamo già usato in precedenza, a cui abbiamo applicato la query:

```
SELECT * FROM penguins where sex !='';
```

Ottenendo così gli esemplari desiderati che sono stati salvati su un nuovo file CSV, chiamato “pinguinipulitisenzaman.csv” ,per essere successivamente caricati su Weka.

Classificazione con l'experimenter

Per effettuare quest'ultima analisi è stato scelto di ricorrere allo strumento «experimenter», disponibile in Weka, che permette di confrontare più tecniche di classificazione contemporaneamente, eseguendo più tecniche in successione sul dataset desiderato, in modo da trovare quelle migliori in tempo ridotto rispetto al metodo visto prima che sfruttava lo strumento «explorer» di Weka considerando singolarmente ogni tecnica.

Abbiamo quindi deciso di applicare le stesse tecniche di classificazione che abbiamo utilizzato in precedenza, ossia: alberi di decisione, regole, vicini più prossimi (ponendo questa volta il numero di vicini da considerare a 4), e quelle basate sul teorema di Bayes e sulle reti neurali artificiali (con il percettrone multilivello).

Inoltre, per quanto riguarda le tecniche basate su alberi di decisione e su regole, è stato deciso di confrontare sia i risultati ottenuti dall'applicazione dei rispettivi metodi lasciando invariati i parametri di default e ponendo invece il limite minimo di esemplari a 10 rispettivamente per un nodo dell'albero di decisione e per una regola.

Risultati classificazione experimenter

Dataset	(1) rules.JR (2) rules (3) trees (4) trees (5) funct (6) bayes (7) lazy.
penguins_pulitisenzaman	(100) 90.11 87.73 88.06 85.26 * 90.48 71.73 * 90.90
	(v/ /*) (0/1/0) (0/1/0) (0/0/1) (0/1/0) (0/0/1) (0/1/0)

Key:

- (1) rules.JRip '-F 3 -N 2.0 -O 2 -S 1' -6589312996832147161
- (2) rules.JRip '-F 3 -N 10.0 -O 2 -S 1' -6589312996832147161
- (3) trees.J48 '-C 0.25 -M 2' -217733168393644444
- (4) trees.J48 '-C 0.25 -M 10' -217733168393644444
- (5) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
- (6) bayes.NaiveBayes '' 5995231201785697655
- (7) lazy.IBk '-K 4 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\\\"\\\" -3080186098777067172

Sopra sono riportati i risultati delle tecniche di classificazione descritte in precedenza considerando il genere come classe, si noti come l'accuratezza maggiore si ottenga, ancora una volta, utilizzando la tecnica dei vicini più prossimi seguita da quella che sfrutta il percepitrone multilivello con un'accuratezza di poco inferiore.

Commento dei risultati

Come è possibile osservare dal lucido precedente, le tecniche basate sul percepitrone multilivello e sui vicini più prossimi risultano essere, ancora una volta, quelle con un'accuratezza maggiore, rispettivamente del 90,48% e 90,90% anche se, come visto in precedenza, non producono un modello di classificazione comprensibile e facilmente interpretabile.

Se cerchiamo quindi una tecnica che restituisca un modello di classificazione di facile interpretazione potremmo essere tentati di sfruttare la tecnica basata su regole, in particolare quella con un numero minimo di 2 esemplari per regola, che ha ottenuto un'accuratezza del 90,11%, tuttavia, come possiamo notare dall'immagine sotto riportata, il modello generato risulta comunque non di facile interpretazione, considerando che è composto da ben cinque regole più una di default.

```
(body_mass_g <= 3700) and (bill_depth_mm <= 18.5) => sex=FEMALE (83.0/2.0)
(bill_depth_mm <= 14.8) and (body_mass_g <= 5200) => sex=FEMALE (51.0/0.0)
(bill_length_mm <= 38.7) and (body_mass_g <= 3725) => sex=FEMALE (11.0/2.0)
(bill_length_mm <= 48.5) and (species = Chinstrap) => sex=FEMALE (8.0/0.0)
(bill_depth_mm <= 15.5) and (body_mass_g <= 5000) and (flipper_length_mm <= 219) => sex=FEMALE (8.0/1.0)
=> sex=MALE (172.0/9.0)
```

Regole derivanti dall'applicazione della tecnica di classificazione basata su regole, imponendo che ogni regola venga rispettata da almeno due esemplari

Commento dei risultati

Se vogliamo quindi ottenere un modello di classificazione più semplice potremo considerare sempre la tecnica basata su regole, ma imporre la condizione che ogni regola debba essere rispettata da almeno 10 esemplari. In questo modo ottengo un modello nettamente più semplice e di facile comprensione, composto da tre regole più una di default, come è possibile osservare nella figura sotto riportata, con un'accuratezza del 87,73%, «perdendo» quindi poco più del 2% di accuratezza rispetto al modello visto in precedenza.

```
(body_mass_g <= 3700) and (bill_depth_mm <= 18.5) => sex=FEMALE (83.0/2.0)
(bill_depth_mm <= 14.8) and (body_mass_g <= 5200) => sex=FEMALE (51.0/0.0)
(body_mass_g <= 3850) and (bill_length_mm <= 39.5) and (bill_depth_mm <= 19.3) and (flipper_length_mm >= 186) => sex=FEMALE (16.0/2
=> sex=MALE (183.0/19.0)
```

Regole derivanti dall'applicazione della tecnica di classificazione basata su regole, imponendo che ogni regola venga rispettata da almeno due esemplari

Commento finale

Abbiamo quindi osservato come la classificazione della specie dei nostri esemplari risulti essere molto più accurata rispetto alla classificazione del genere. Questa differenza può essere dovuta in parte anche all'attributo riguardante l'isola di provenienza dell'esemplare, poiché come avevamo già evidenziato nella fase di data understanding, solo i pinguini della specie Adelie sono presenti su tutte le isole dell'arcipelago considerate, mentre le altre due specie sono presenti solo su un'isola, questo probabilmente ha permesso alle varie tecniche di classificazione di discriminare in maniera più corretta gli esemplari rispetto alla specie che al genere.

Infatti gli esemplari di genere maschile e femminile sono distribuiti in maniera omogenea sia tra le specie che tra le isole (cosa non troppo sorprendente, considerato che in natura è raro trovare popolazioni nelle quali uno dei due generi è preponderante rispetto all'altro).

Conclusione del nostro studio

Riassumiamo quindi i risultati delle nostre analisi e riportiamo le conclusioni a cui siamo giunti

Conclusioni

Nel nostro lavoro, in particolare attraverso il clustering, abbiamo osservato come, almeno per gli attributi presi in analisi nel nostro dataset, il gruppo di pinguini appartenenti alla specie «Gentoo» risulti possedere caratteristiche che lo distinguono in maniera abbastanza netta dalle altre due specie, che invece risultano avere differenze meno marcate tra loro.

Abbiamo quindi individuato come miglior algoritmo di clustering quello gerarchico agglomerativo con metodo del legame medio per il calcolo della distanza tra cluster, che ha raggruppato nel cluster incorrecto appena il 3,22% degli esemplari in esame, il cui dendogramma risultante ha inoltre permesso di evidenziare la differenza presente tra la specie «Gentoo» e le altre due specie analizzando la distanza di fusione tra i cluster.

Conclusioni

Per quanto riguarda invece la classificazione, abbiamo osservato come le tecniche con accuratezza maggiore, sia che si consideri come classe la specie o il genere degli esemplari, siano risultate essere quella basata sulle reti neurali artificiali, in particolare con il percepitrone multilivello, e quella basata sui vicini più prossimi.

Tuttavia, se desideriamo un modello relativamente semplice e comprensibile, è stato evidenziato come in entrambi i casi la tecnica basata sulle regole restituiscia due insiemi di regole relativamente semplici da comprendere, anche se tali tecniche presentano un'accuratezza inferiore rispetto alle altre due tecniche citate in precedenza.

Possiamo quindi affermare di aver trovato varie tecniche efficaci per classificare esemplari anche estranei al nostro insieme di partenza, permettendo quindi di distinguere con sufficiente certezza gli individui per genere o per specie, cosa che in particolare per il genere non è così scontata, visto che lo scopo del lavoro da cui proviene il dataset era proprio lo studio del dismorfismo sessuale in ognuna delle tre specie di pinguino.

Considerazioni finali

Questa analisi ha permesso di caratterizzare e di comprendere la composizione e le caratteristiche delle specie di pinguino presenti nell'arcipelago di Palmer e quanto esse differiscano tra loro, arrivando poi ad ottenere delle tecniche di classificazione che abbiamo visto essere molto accurate.

Questo tipo di analisi, svolte su un insieme comunque abbastanza ampio di esemplari, potrebbe rivelarsi fondamentale in futuro a causa soprattutto del cambiamento climatico, che, seppur nessuna delle tre specie risulti attualmente a rischio imminente di estinzione, potrebbe in futuro causare seri danni alla conservazione delle specie che abbiamo studiato, in quanto, con lo scioglimento dei ghiacci ed il conseguente aumento del livello del mare, il loro habitat sta venendo consumato in maniera incredibilmente rapida, tanto da spingere l'unione internazionale per la conservazione della natura a catalogare le tre specie come «prossime alla minaccia».

È quindi di fondamentale importanza riuscire a studiare e caratterizzare il più possibile queste specie prima che i loro esemplari diminuiscano drasticamente, per riuscire ad applicare strategie di conservazione e di ripopolamento che risultino mirate ed efficaci, evitando che questi magnifici esemplari vadano incontro ad un terribile destino.

FINE



Articolo da cui è stato ricavato il dataset:

<https://doi.org/10.1371/journal.pone.0090081>