Projektová dokumentácia UPA

Ondrej Valo, xvaloo00@stud.fit.vutbr.cz
Radoslav Páleník, xpalen05@stud.fit.vutbr.cz
Karel Fritz, xfritz00@stud.fit.vutbr.cz

zima 2022

1 Exploratívna analýza

1.1 Atributy datové sady

Dataset obsahuje celkovo 344 záznamov s atribútmi, ktoré jemožné rozdeliť na 2 podmnožiny:

1.1.1 Atribúty odberu

- **studyName** [*string*]
  Štúdia z ktorej pochádza odber. V datasete sú záznamy zo štúdií *PAL0708*, *PAL0809* a *PAL0910*.
- **Sample Number** [*int*]
  Číslo vzorky odberu. Jednotlivé vzorky sú číslované chronologicky v poradí odberov.
- **Region** [*string*]
  Región merania. Všetky zaznamenané odbery prebehli v regióne Anvers.
- **Island** [*string*]
  Ostrov Palmerského súostrovia. Definuje ostrov, na ktorom bol daný odber uskutočnený. V datasete sa nachádzajú celkovo 3 ostrovy: *Torgersen*, *Biscoe* a *Dream*. Najpočetnejšie zastúpenie v meraniach má ostrov *Biscoe* s celkovým podielom 49%.

1.1.2 Atribúty sledovaného tučniaka

- **Species** [*string*]
  Druh tučniaka. Dataset sleduje merania na troch druhoch tučniakov: *Adelie*, *Chinstrap* a *Gentoo*. Najpočetnejším druhovým zástupcom v záznamoch je *Adelie* s podielom 44% vo všetkých výskytoch.
- **Stage** [*string*]
  Reprodukčné štádium počas odberu. Pre tento atribút sa v datasete objavuje len 1 hodnota, „Adult, 1 Egg Stage“. Dataset teda uvažuje jedno vajce v každej znáške.
- **Individual ID** [*string*]
  Unikátny identifikátor študovaného tučniaka. Celkovo sa v datasete nachádza 190 unikátnych jedincov.
- **Clutch Completion** [*bool*]
  Dokončenie znášky. Počas meraní sa sleduje, či v čase merania bol splodený nový jedinec. V datasete sa nachádza 89.5% záznamov s hodnotou *true*.
- **Date Egg** [*date*]
  Dátum odberu vzorky. Zaznamenané odbery prebehli v rozmedzí od 9.11.2007 do 1.12.2009.
- **Sex** [*string*]
  Pohlavie tučniaka, v datasete sa uvádzajú 2 rôzne hodnoty: *FEMALE* a *MALE*.

- **Culmen Length (mm)** [*float*]
  Dĺžka vrchnej časti zobáku (v *mm*)

Minimum	Priemer	Maximum
32.1	43.92	59.6

- **Culmen Depth (mm)** [*float*]
  Hĺbka vrchnej časti zobáku (v *mm*)

Minimum	Priemer	Maximum
13.1	17.15	21.5

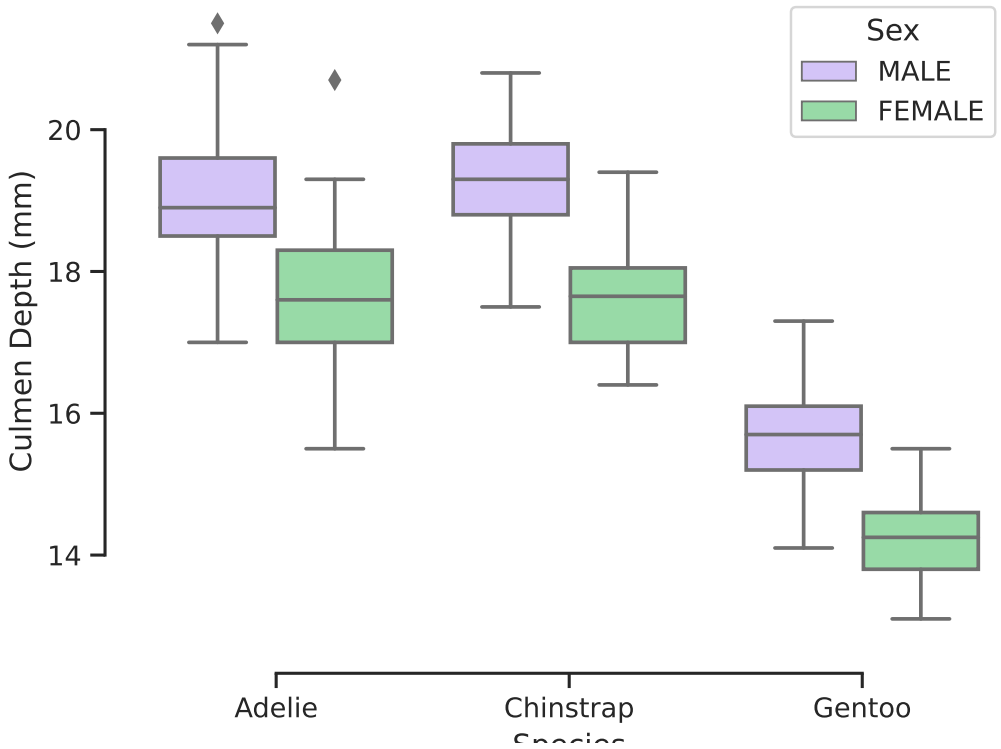

Figure 1: Distribúcia hĺbky vrchnej časti zobáku

- **Flipper Length (mm)** [*float*]
  Dĺžka plutvy (v *mm*)

Minimum	Priemer	Maximum
172	200.92	231

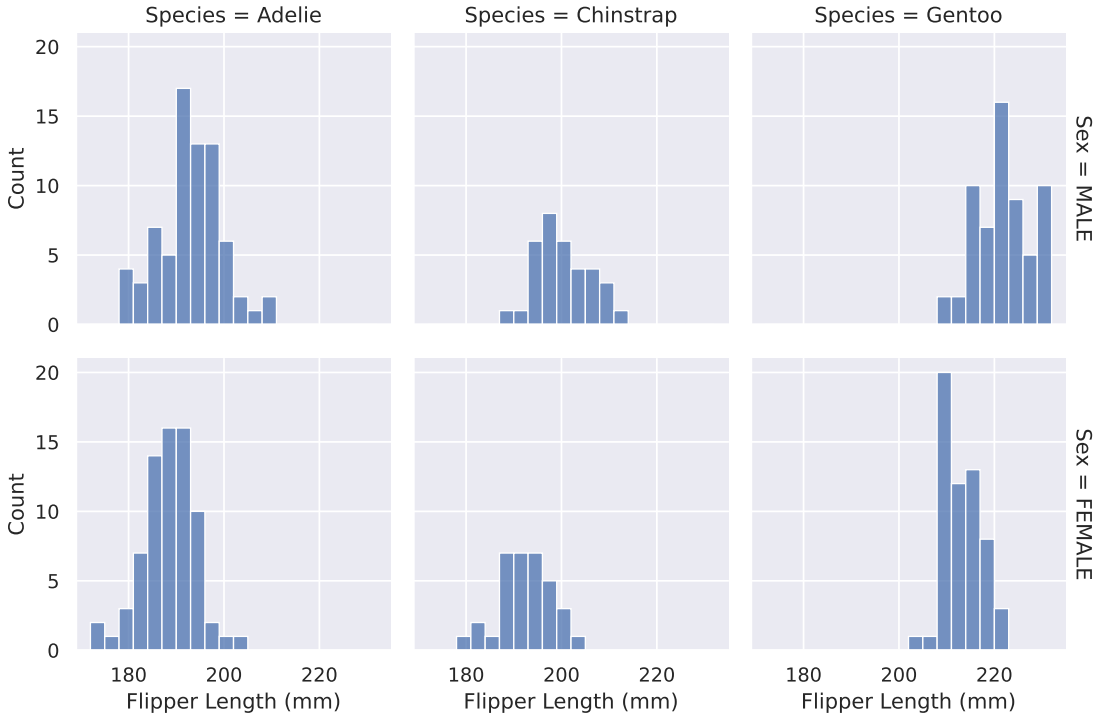

Figure 2: Distribúcia dĺžky plutiev

- **Delta 15 N (o/oo)** [*float*]
  Miera pomeru stabilných izotopov 15N:14N

Minimum	Priemer	Maximum
7.63	8.73	10.03

- **Delta 13 C (o/oo)** [*float*]
  Miera pomeru stabilných izotopov 13C:12C

Minimum	Priemer	Maximum
-27.02	-25.69	-23.79

- **Body Mass (g)** [*float*]
  Telesná masa tučniaka (v g)

Minimum	Priemer	Maximum
2700	4201.75	6300

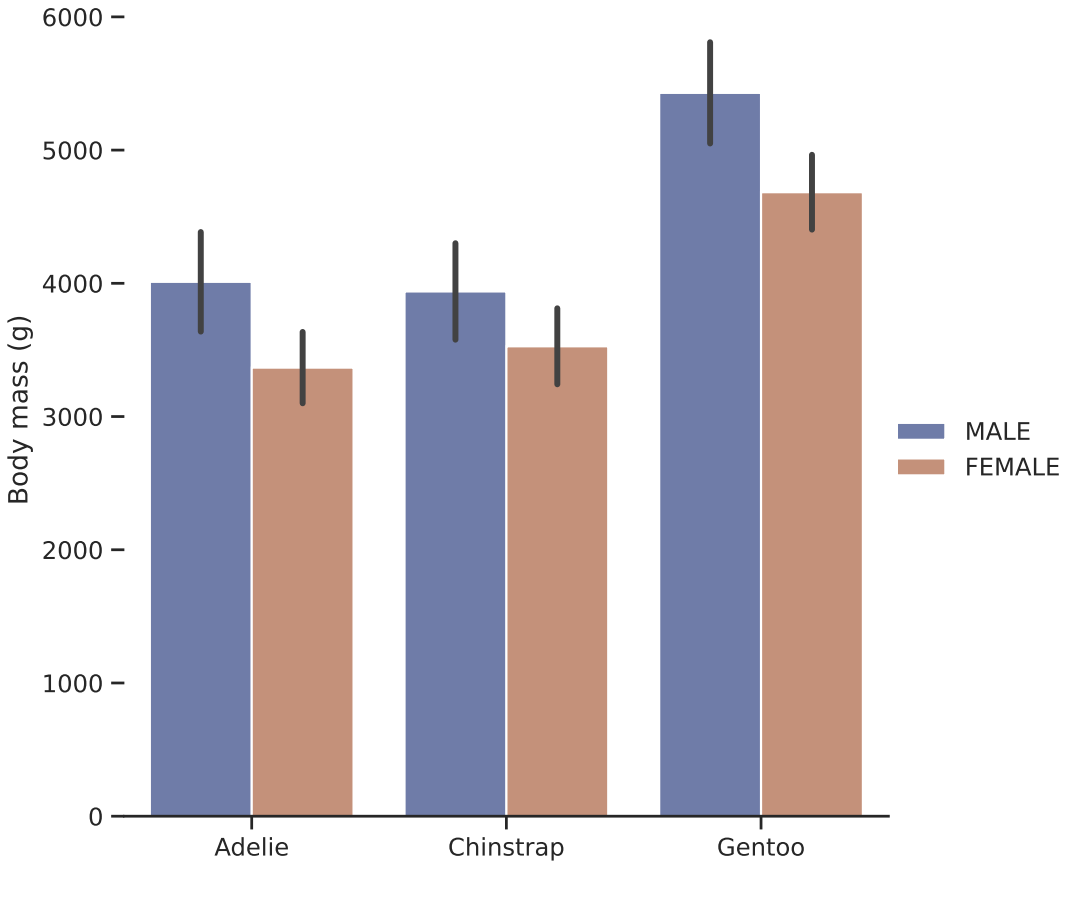

Figure 3: Distribúcia telesnej masy tučniakov

- **Comments** [*string*]
  Komentár k meraniu určený na obsiahnutie iných relevantných informácií k dátam.

1.2 Analýza chýbajúcich hodnôt

V databáze sa celkovo nachádza 12 záznamov s 2 a viac chýbajúcimi atribútmi. Majoritu chýbajúcich atribútov tvorí atribút *Comments* (chýba v 318 záznamoch). Tento atribút sa totiž využíva najmä ako poznámka pri nekompletných záznamoch, príp. pri záznamoch jedincov u ktorých nebola počas skúmania zaznamenaná znáška. Okrem komentárov je početnosť chýbajúcich atribútov nasledovná:

Atribút	Počet	Najčastejšie odôvodnenie
Delta 15 N (o/oo)	14	Nedostatok krvi na testy
Delta 13 C (o/oo)	13	Nedostatok krvi na testy
Sex	11	Nedostatok krvi na testy
Culmen Length (mm)	2	Jedinec nebol meraný
Culmen Depth (mm)	2	Jedinec nebol meraný
Flipper Length (mm)	2	Jedinec nebol meraný
Body Mass (g)	2	Jedinec nebol meraný

1.3 Korelačná analýza numerických atribútov

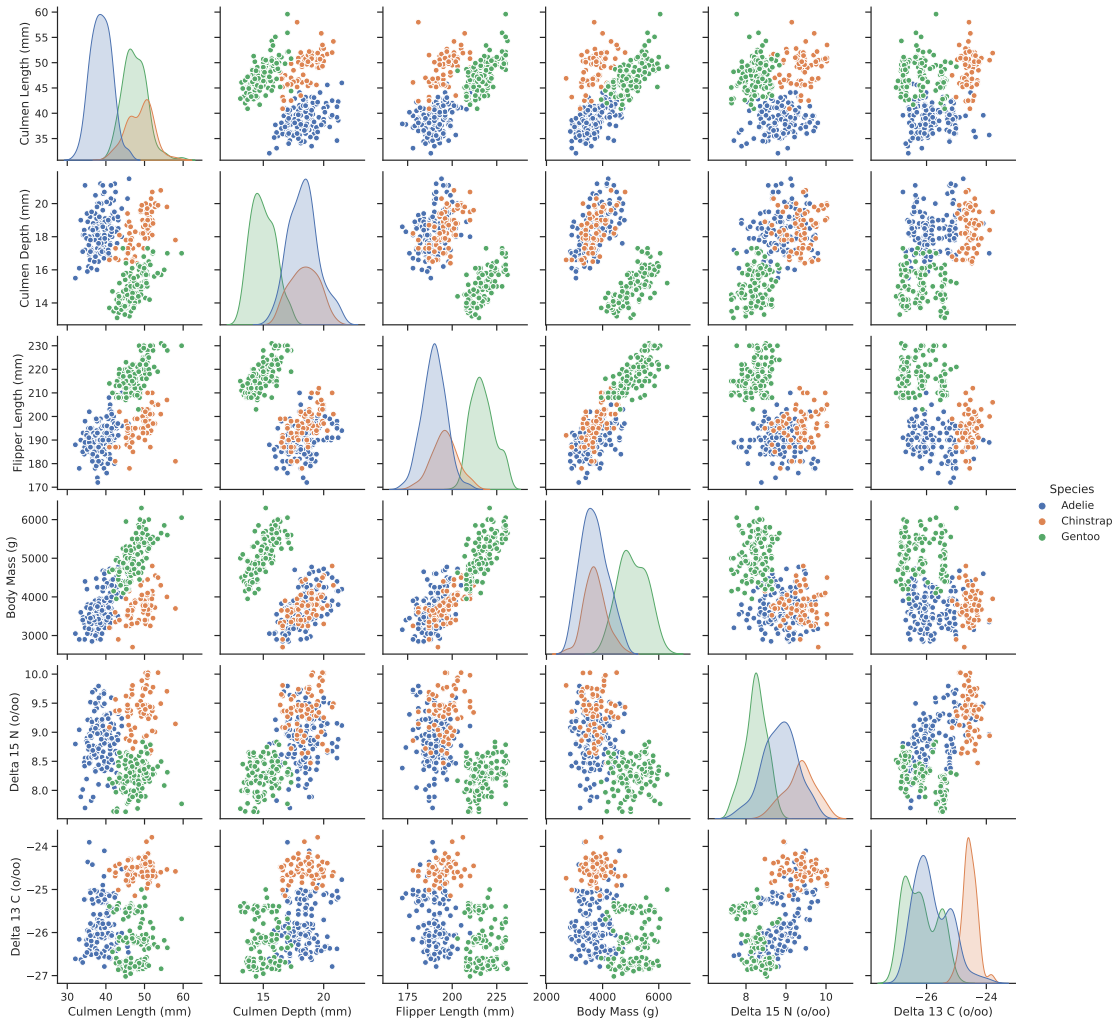

Figure 4: Korelácia dát medzi všetkými numerickými atribútami

Z grafov 4 môžeme pozorovať nenulovú koreláciu skoro pre všetky numerické atribúty danej dátovej sady. Aj keď je možno vidieť vyšiu koreláciu medzi niektorými atribútami ako je dĺžka krídiel a dĺžka hrany zobáka, oproti

nižším ako je hmotnosť s hĺbkou zobáka pre druhy Adelie a Chinstrap. Koreláciu v nie veľmi očakávaných atribútoch si môžeme vysvetliť, vysokým rozdielom v daných atribútov medzi druhmi ča nám ukazujú grafy po diagonále obrázku.

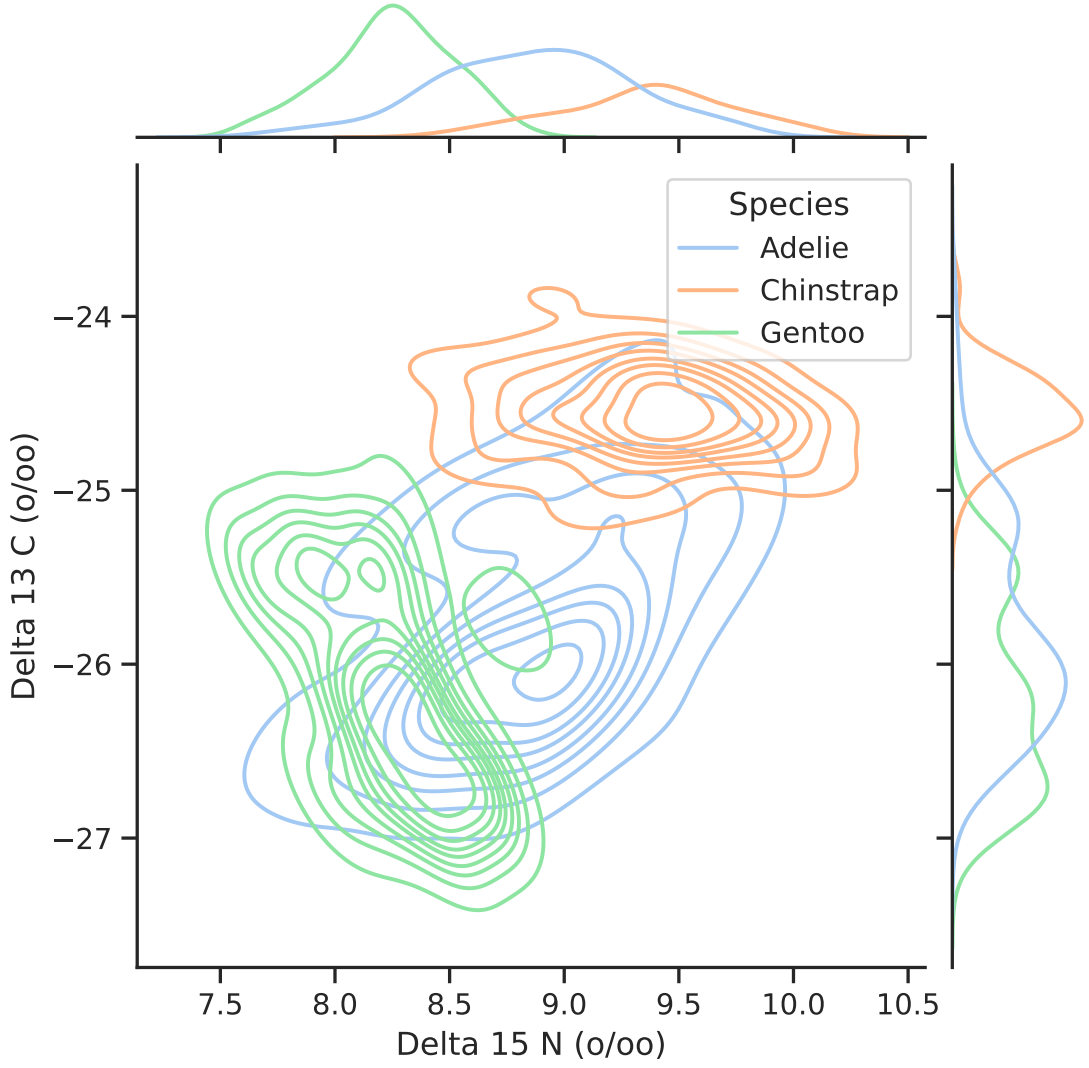

Figure 5: Korelácia izotopov v krvi

Z grafu 5 môžeme vyčítať nižší počet izotopov Typu C a N pre tučniaky druhu Gentoo, a vyšší pre Chinstrap. Tučniaky Adelie majú väčší rozsah počtu izotopov.

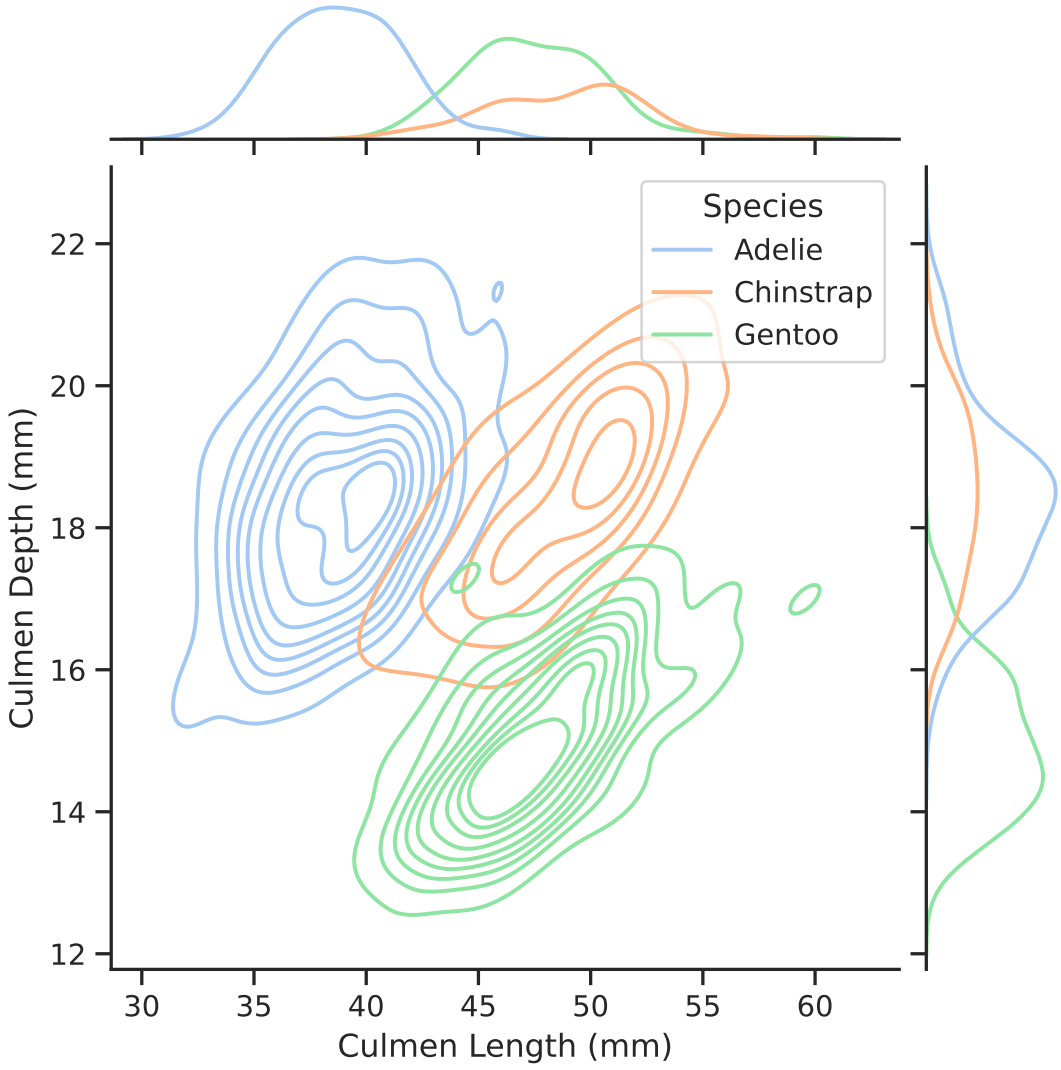

Figure 6: Korelácia vrchnej časti zobáku

Z grafu 6 môžeme vyčítať že tučniaky Adelie majú nižšiu dĺžku ale väčšiu hĺbku hrany zobáka, Gentoo sú presný opak a Chinstrap majú rovnomerne dlhú a hlbokú hranu zobáku.

2 Príprava dátovej sady pre dolovacie algoritmy

Prvým dôležitým krokom je odstrániť nerelevantné informácie. Jedná sa o informácie, ktoré z logického hľadiska neprispievajú k správnej klasifikácii. Ako nerelevantné môžeme považovať informácie zo stĺpcov „studyName, Sample Number, Region, Stage, Individual ID, Clutch Completion, Date Egg, Comments“. Keďže ani jeden z týchto stĺpcov nie je informačne pre klasifikáciu dôležitý, sú tieto stĺpce odstránené.

2.1 Kategorická príprava

Pri týchto dátach boli pre chýbajúce hodnoty odstránené celé riadky(záznamy). Odstránenie hodnôt bolo prevedené pomocou *DataFrame.dropna()* z knižnice *pandas*. Ďalej sa tu odľahlé hodnoty nevyskytovali, dataset bol vysokej kvality s menším počtom záznamov, mierné vyhladenie hodnôt prebehlo vďaka metóde *qcut()* opäť z knižnice *pandas*. Táto metóda prevedie kvantilovú diskretizáciu, teda rozdelí dáta do n košov tak, aby v každom koši bol rovnaký počet záznamov. Pri uvedenej implementácii sa používa 8 košov. Tým je príprava dát pre klasifikátor pracujúci s kategorickými vstupmi dokončená.

2.2 Numerická príprava

U chýbajúcich hodnôt tu bola vykonaná interpolácia, aj pomocou metódy z knižnice *pandas*, konkrétne *DataFrame.interpolate*
Keďže v dátach neboli žiadne významné odľahlé hodnoty, prešlo sa ihneď k numerifikácii. Metóda *get_ dummies()* funguje na princípe algoritmu one-hot encoding, ktoré vytvorí z kategorických hodnôt numerické. Naviac zachová rovnakú dôležitosť vďaka zakódovaniu iba pomocou 0 a 1, teda pokiaľ bol tučniak samec, budú hodnoty pre stĺpce priradené následovne: Gender-male 1 a Gender-female 0.

Normalizácia hodnôt prebehla nasledovne: Pre všetky stĺpce sa nahradia ich hodnoty novou hodnotou získanou ako podiel hodnoty v riadku daného stĺpcu s maximálnou hodnotou daného stĺpcu. Tím je dataset pripravený pre klasifikátor využívajúce numerické vstupy.