VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

UPA

Analýza a příprava dat

Projekt, 2. část

Jan Klhůfek (xklhuf01)
Andrea Chimenti (xchime00)
Jiří Václavič (xvacla31)

22. listopadu 2022

1 Úvod

Cílem projektu bylo nad zvolenou datovou sadou nejprve provést explorační analýzu za účelem bližšího pochopení dat a následně upravit sadu do dvou variant vhodných pro dolovací algoritmy.

K bližšímu prozkoumání byla vybrána datová sada tučňáků vzhledem ke své popularitě jakožto testovací datový soubor užívaný pro různé statistické klasifikační techniky. Sada je volně ke stažení zde: dataset. Dolovací úlohou je pak klasifikace druhů tučňáků na základě ostatních atributů.

Analýza, vizualizace a úprava dat byly provedeny s využitím modulů *pandas*, *seaborn*, *mathplotlib* programovacího jazyka Python 3.10.

2 Datová sada tučňáků

Informace o tučňácích byly sesbírány z 3 ostrovů ze Souostroví Palmer (Antarktida) a obsahují údaje o jedincích patřících do 3 různých tamních druhů. Datový soubor obsahuje celkem 344 záznamů a 17 atributů, z nichž 10 je kategorických a 7 numerických. Kategorické atributy zobrazuje obrázek 1 a numerické viz obrázek 2. Bližší popis jednotlivých atributů je dostupný na: palmerpenguins.

studyName	Sample Number	Species	Region	Island	Stage	Individual ID	Clutch Completion	Date Egg	Sex	Comment

Obrázek 1: Výčet kategorických atributů datové sady.

Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Delta 15 N (o/oo)	Delta 13 C (o/oo)

Obrázek 2: Výčet numerických atributů datové sady.

3 Explorativní analýza

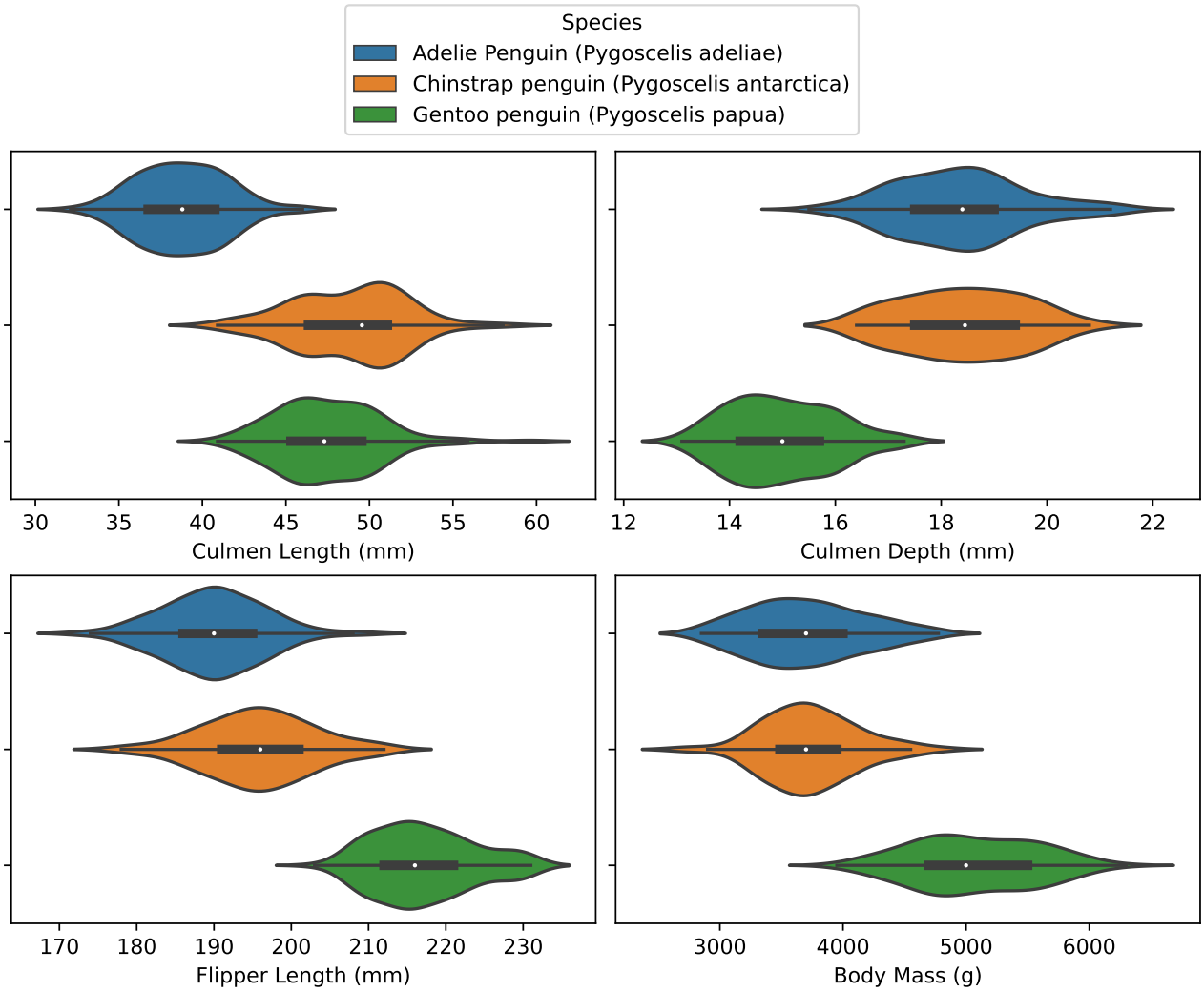
3.1 Průzkum atributů

	studyName	**Sample Number**	**Species**	**Region**	**Island**	**Stage**	**Individual ID**
Počet unikátních hodnot	3	152	3	1	3	1	190
Nejčastější hodnota	PAL0910	1	Adelie Penguin	Anvers	Biscoe	Adult	N61A2

Tabulka 1: Počet unikátních hodnot nejčastější výskyt dle atributů

3.2 Rozložení hodnot

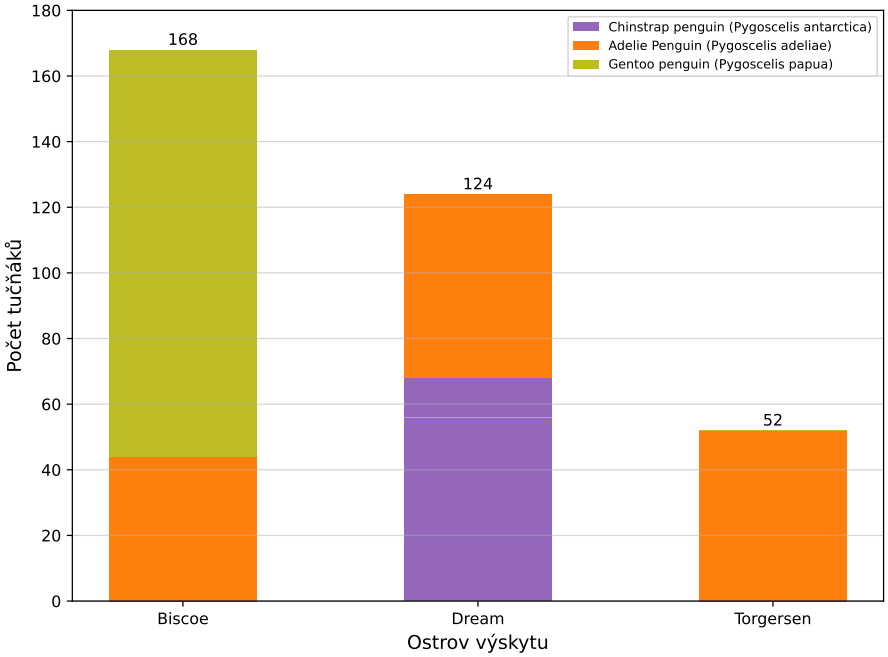
Pro provedení průzkumu rozložení atributů byly použito 5 různých grafů. Vytvořili jsme grafy rozložení nad potenciálně zajímavými údaji. Zkoumali jsme rozložení hodnot tělesných vlastností tučňáků podle druhu, které jsme zobrazili v houslových grafech. Z grafu na obrázku č. 3 lze například vyčíst, že druh *Gentoo penguin (Pygoscelis papua)* dosahuje jako celek největší hmotnosti.



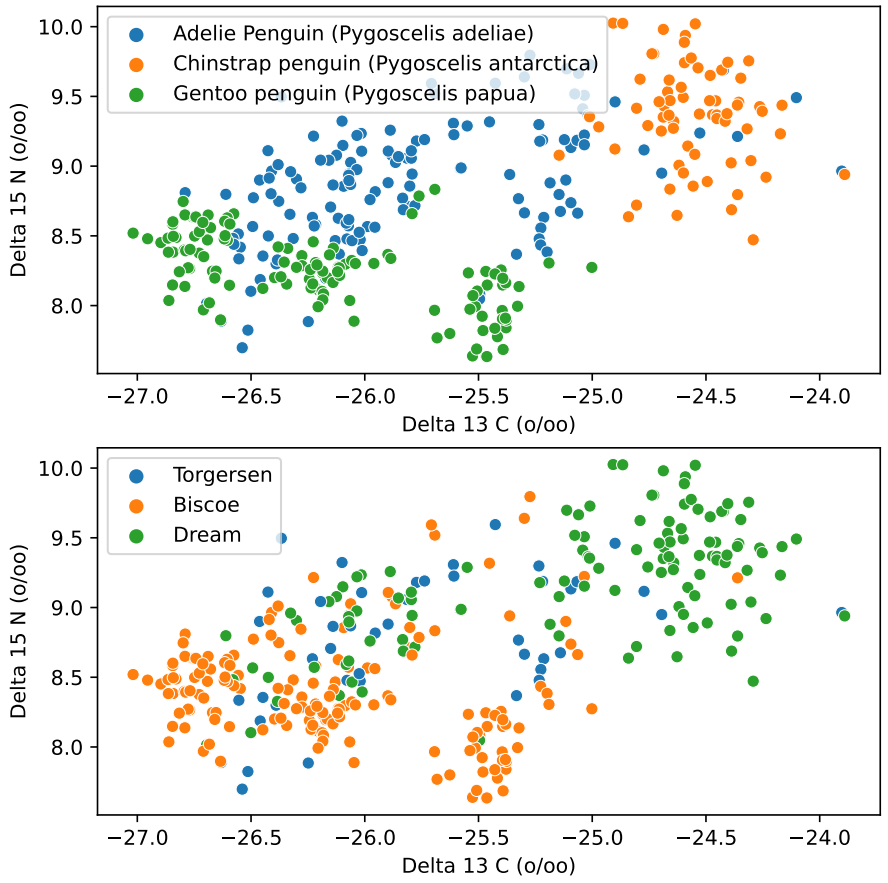
Obrázek 3: Houslový graf rozložení hodnot tělesných vlastností tučňáků podle druhu

Dále jsme zkoumali zastoupení tučňáků na jednotlivých ostrovech. Na sloupcovém grafu na obrázku č. 4 lze vidět, že největší množství tučňáků se náchází na ostrově *Biscoe* a dominatním druhem je zde *Gentoo penguin (Pygoscelis papua)*.

Následně jsme zkoumali hodnoty *Delta15N* a *Delta13C* (hodnoty izotopu uhlíku a dusíku nacházejících se v krvi, peří a kostech tučňáků), v závislosti na ostrově a druhu. Z bodových grafů na obrázku č. 5 můžeme vypozorovat, že významnější shluky tvoří pouze určení podle druhu. Konkrétně druhy Chinstrap a Gentoo. Ostrovy se nezdají mít příliš velký vliv na poměr rozložení izotopů dusíku ani uhlíku.

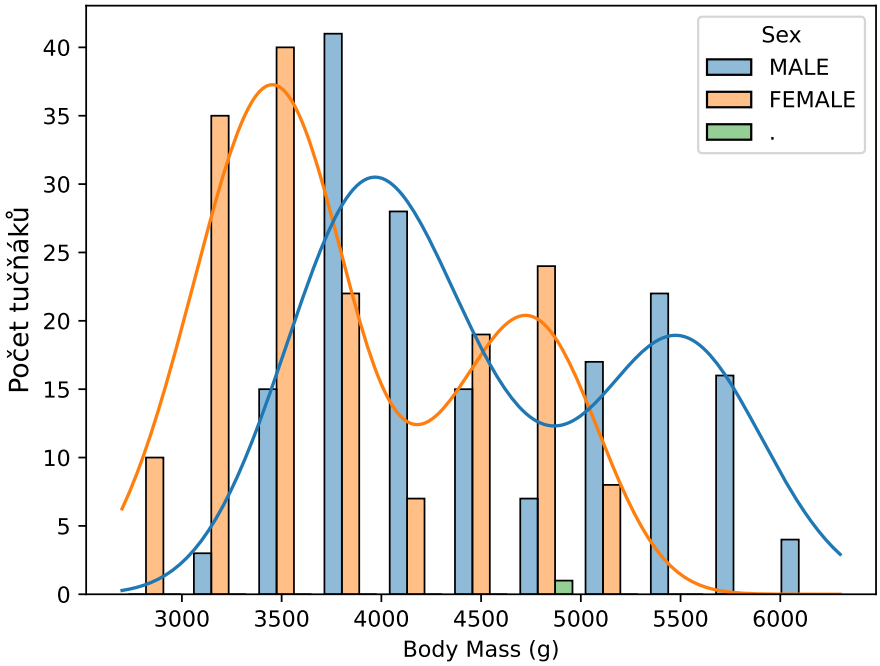


Obrázek 4: Sloupcový graf zastoupení tučňáků dle druhu na jednotlivých ostrovech



Obrázek 5: Hodnoty Delta15N a Delta13C v závislosti na ostrově a druhu

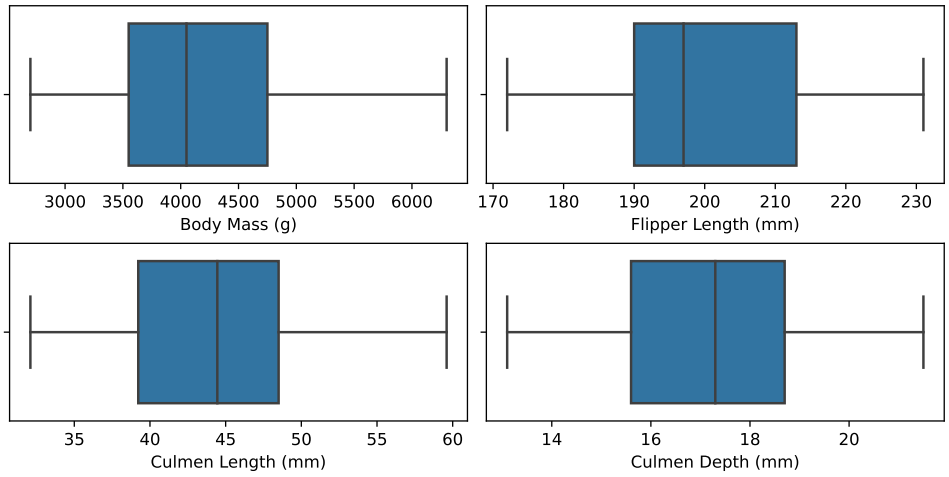
Na dalším grafu na obrázku č. 6 nás zajímala hmotnost tučňáků v závislosti na pohlaví. Z grafu je viditelné, že největší počet tučňáků obou pohlaví se koncentruje u dvou hmotnostních bodů. Nejvíce samic má hmotnost okolo 3500 g a samců okolo 4000 g. Druhá nejčastější hmotnost samic je okolo 4700 g a samců 5500 g. Z grafu je navíc identifikovatelná i nevalidní hodnota pohlaví ve vstupních datech.



Obrázek 6: Histogram hmotnosti tučňáků v závislosti na jejich pohlaví

3.3 Odlehlé hodnoty

V první fázi se podíváme na krabicové grafy na obrázku č.7 pro zvolené atributy. Z grafů vidíme, že žádné odlehlé hodnoty nebyly nalezeny. Je však vhodné si tuto skutečnost ověřit i numericky. Proto dále vypočítáme z-score, které pro každý vzorek určí, jak daleko od střední hodnoty se nachází. Hodnota je udáná ve směrodatných odchylkách. Za odlehlé hodnoty budeme považovat data, která jsou alespoň 3 σ od μ. Jelikož takové hodnoty neexistují, tak je v tabulce č. 3 ukazka prahu 1 σ.



Obrázek 7: Krabicové grafy pro zvolené atributy

	studyName	Sample Number	Species	Region	Island	Stage	Individual ID	Clutch Completion
Počet unikátních hodnot	3	152	3	1	3	1	190	2
Nejčastější hodnota	PAL0910	1	Adelie Penguin	Anvers	Biscoe	Adult	N61A2	Yes

Tabulka 2: Z-score pro vybrané atributy s prahem 1 σ

	Date Egg	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex	Delta 15 N (o/oo)	Delta 13 C (o/oo)	Comments
Počet unikátních hodnot	50	165	81	56	95	4	331	332	8
Nejčastější hodnota	11/27/07	41.1	17	190	3800	MALE	8.95	-24.69	Nest never observed with full clutch.

Tabulka 3: Z-score pro vybrané atributy s prahem 1 σ

3.4 Chybějící hodnoty

Z analýzy chybějících hodnot bylo zjištěno, že alespoň jedna hodnota chybí u 8 atributů. Celkový počet všech chybějících je roven 363. Celkový počet chybějících položek u jednotlivých atributů je zanesen do tabulky č. 4 níže.

Culmen Length (mm)	2
Culmen Depth (mm)	2
Flipper Length (mm)	2
Body Mass (g)	2
Sex	10
Delta 15 N (o/oo)	14
Delta 13 C (o/oo)	13
Comments	318
Součet všech chybějících hodnot:	**363**

Tabulka 4: Chybějící hodnoty jednotlivých atributů

3.5 Korelační analýza numerických atributů

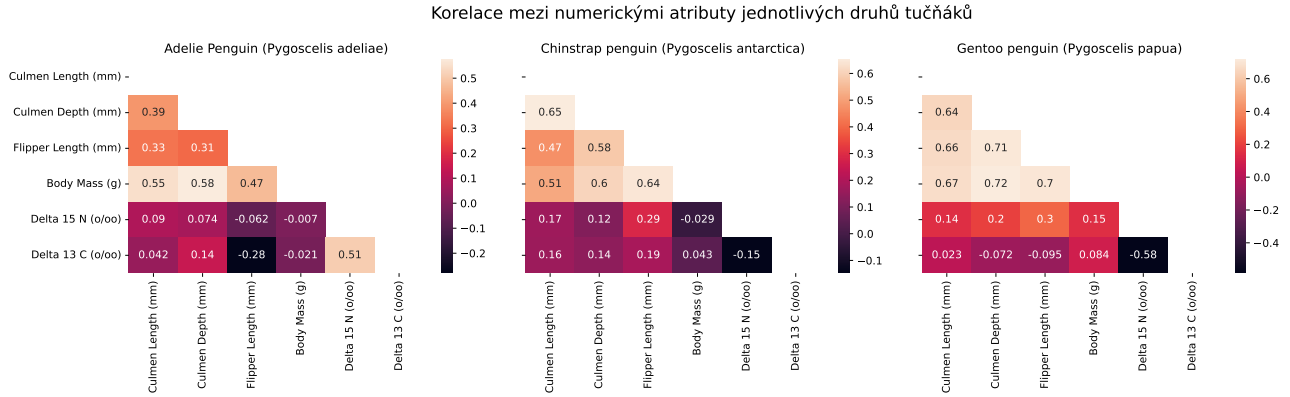
Z datasetu jsme získali data pouze numerických atributů. K zobrazení korelací, tedy zjištění zda-li mezi atributy existují závislosti jsme použili teplotní mapu (”heatmap”). Čím větší je korelace tím výraznější je zbarvení čtverce. Z tabulky č. 5 níže je patrné, že zde existuje spousta středně silných zavislostí a dokonce i jedna silná přesahující hodnotu 0.7.

Silná závislost je mezi atributy *Body Mass (g)* a *Flipper Length (mm)*, jejichž korelační koeficient je 0.87. Mezi další středně silné závislosti patří například ”Flipper Length (mm)”a ”Culmen Length (mm)”s korelačním koeficientem 0.66 a atributy ”Delta 15 N (o/oo)”and ”Culmen Depth (mm) s korelačním koeficientem 0.61.

Za bližší prozkoumání stojí i korelační analýza mezi stejnými atributy u dílčích druhů tučňáků, viz hodnoty teplotních map na grafu 8.

	Body Mass	**Culmen Length**	**Culmen Depth**	**Flipper Length**	**Delta 15 N**	**Delta 13 C**
Body Mass (g)	1.00	0.60	-0.47	0.87	-0.54	-0.37
Culmen Length (mm)	0.60	1.00	-0.24	0.66	-0.06	0.19
Culmen Depth (mm)	0.47	-0.24	1.00	-0.58	0.61	0.43
Flipper Length (mm)	0.87	0.66	-0.58	1.00	-0.51	-0.38
Delta 15 N (o/oo)	-0.54	-0.06	0.61	-0.51	1.00	0.57
Delta 13 C (o/oo)	-0.37	0.19	0.43	-0.38	0.57	1.00

Tabulka 5: Korelace jednotlivých atributů nad všemi tučňáky.



Obrázek 8: Teplotní mapa zobrazující korelace mezi numerickými atributy u jednotlivých druhů tučňáků.

4 Úprava datové sady

4.1 Odstranění irelevantních atributů

Z datové sady bylo odstraněno celkem 9 původních atributů nepotřebných pro úlohu klasifikace. Výběr redukované podmnožiny atributů byl proveden na základě informací zjištěných explorační analýzou. Konkrétně byly odstraněny atributy sloužící jako identifikátory (*Sample Number, Individual ID*), atributy s 1 unikátní hodnotou (*Region, Stage*) nebo atributy, které pro klasifikaci nejsou významné či by mohly snížit její přesnost (*studyName, Island, Clutch Completion, Date Egg, Comment*). Vybraná podmnožina atributů viz 9.

Species	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex	Delta 15 N (o/oo)	Delta 13 C (o/oo)

Obrázek 9: Podmnožina atributů datové sady vybraná pro použití dolovacími algoritmy.

4.2 Odstranění chybějících hodnot

Po redukci dimenzionality datové sady je z tabulky 4 patrné, že chybí hodnoty u všech numerických atributů a také u atributu pohlaví. Analýzou bylo zjištěno, že u 2 záznamů chybí všechny zmíněné atributy, proto byly záznamy z datového souboru odstraněny. Pro jednu variantu datové sady byl pro nahrazení chybějících numerických hodnot použit průměr z ostatních hodnot daného atributu, zatímco pro druhou variantu se hodnoty doplnily mediánem. Chybějící hodnoty pohlaví byly doplněny na základě nejmenší vzdálenosti mezi hmotností tučňáka od průměru hmotností pro první variantu, respektive mediánu hmotností pro druhou variantu, pro patřičný druh tučňáka a jednotlivá pohlaví.

4.3 Odstranění odlehlých hodnot

Žádné numerické odlehlé hodnoty nebyly nalezeny. Jedinou patrnou chybou ve vstupních datech je nevalidní hodnota pohlaví u jednoho záznamu, kde se vyskytla '.' místo 'MALE/FEMALE'. Hodnota byla nejprve nahrazena prázdnou hodnotou a následně doplněna odhadem, viz předchozí podsekce.

4.4 Diskretizace numerických atributů (varianta 1)

Pro všech 6 numerických atributů se provedla diskretizace s užitím ekvifrekvenčních intervalů. Původní hodnoty byly rozděleny do 4 intervalů majících přibližně stejnou četnost hodnot. Původní hodnoty pak byly nahrazeny intervalem, do kterého spadají.

4.5 Transformace kategorických atributů, normalizace numerických atributů (varianta 2)

Hodnoty numerických atributů byly normalizovány s použitím min-max normalizace, jelikož jednotlivé atributy mají různý rozsah, ale nemají normální rozdělení. Při transformaci kategorických atributů *Species, Sex* došlo k nárůstu dimenzionality, jelikož se pro každou hodnotu atributů přidal vlastní sloupec a sloupce původní byly odstraněny.

4.6 Upravené varianty datových souborů

Výsledné upravené varianty datových sad byly exportovány do odpovídajících souborů ve formátu csv.