Sprawozdanie z laboratorium: Systemy Uczące się

Studium przypadku

11 czerwca 2023

Prowadzący: dr hab. inż. Maciej Komosiński, prof. PP

Autor: **Agnieszka Klimek** inf145302 SI agnieszka.r.klimek@student.put.poznan.pl

Zajęcia piątkowe, 11:45.

Użyte narzędzia

- `scikit-learn`: algorytmy uczenia maszynowego, metryki,
- `imblearn`: metryka g-mean,
- `pandas`: przechowywanie oraz przekształcanie danych,
- `matplotlib, seaborn`: wizualizacja,
- `scipy`: związki między atrybutami.

1 Opis zbioru danych

Omawiany zbiór danych to Horse Colic Dataset [1]. Zawiera on dane medyczne koni chorujących na kolkę - schorzenie przewodu pokarmowego, które może stanowić bezpośrednie zagrożenie życia. Jest to najczęstsza przycyzna nagłych interwencji weterynaryjnych [3]. Śmiertelność nie jest jednoznacznie określona, zależy od wielu czynników, ogólna przeżywalność oscyluje w okolicach 70% [2, 4]. Zbiór treningowy zawiera 299 obserwacji, a testowy 89. Jest to problem klasyfikacji, celem jest przewidzenie wartości atrybutu „outcome", który może przyjmować jedną z 3 wartości, w zależności od losu zwierzęcia: „lived", „died", „euthanized". W zbiorze obecnych jest, nie licząc atrybutu decyzyjnego, 27 atrybutów kategorycznych oraz numerycznych:

- „surgery" – czy zwierzę było operowane, wartości „Yes", „No";
- „age" – wiek zwierzęcia, wartości „young" (poniżej 6 miesięcy), „adult" (co najmniej 6 miesięcy);
- „hospital number" – wartość numeryczna, określa id pacjenta, przy czym możliwe jest wystąpienie duplikatów, jeżeli koń był leczony więcej niż raz;
- „rectal temperature" – temperatura zwierzęcia w stopniach Celsjusza, mierzona w odbycie, u zdrowego osobnika powinna wynosić 37,8°C;
- „pulse" – atrybut numeryczny, liczba uderzenia serca na minutę, u zdrowego, dorosłego osobnika normą jest 30–40 uderzeń (u koni sportowych może być niższa, nawet 20–25 uderzeń);
- „respiratory rate" – wartość numeryczna, częstość oddechów na minutę, norma powinna mieścić się między 8 a 10;
- „temperature of extremities" – subiektywne wskazanie krążenia obwodowego, możliwymi wartościami są: „normal", „warm", „cool", „cold", przy czym chłodne lub zimne kończyny wskazują na możliwy wstrząs;
- „peripheral pulse" – puls obwodowy, możliwe wartości to: „normal", „increased", „reduced", „absent", normalny lub podwyższony wskazuje na odpowiednie krążenie, podczas gdy obniżony lub nieobecny może sugerować słabą perfuzję;
- „mucous membranes" – kolor błon śluzowych, możliwe wartości to: „normal pink", „bright pink", „pale pink", „pale cyanotic", „bright red", „dark cyanotic"; sine wskazują na poważną niewydolność krążenia, a czerwone świadczą o możliwej infekcji (posocznicy);
- „capillary refill time" – czas napełniania się naczyń włosowatych, możliwe wartości to: „$<$ 3 seconds", „$\geqslant$ 3 seconds", „3 seconds"; dłuższe napełnianie się naczyń sugeruje gorsze krążenie;
- „pain" – subiektywna ocena poziomu bólu odczuwanego przez zwierzę, możliwe wartości to: „alert, no pain", „depressed", „intermittent mild pain", „intermittent severe pain", „continuous severe pain";
- „peristalsis" - aktywność jelit konia, możliwe wartości to: „hypermotile", „normal", „hypomotile", „absent"; aktywność jelit spada, gdy jelita są bardziej rozdęte lub poziom toksyn wzrasta;
- „abdominal distension" – wzdęcie brzucha, możliwe wartości: „none", „slight", „moderate", „severe";

- „nasogastric tube” – odnosi się do gazów wydostających się przez rurkę zgłębnika nosowo–gardłowego, możliwe wartości to: „none”, „slight”, „significant”;
- „nasogastric reflux” – określa ilość zarzucanej treści żołądkowej, możliwe wartości to: „none”, „$<$ 1 liter”, „$>$ 1 liter”, im większa ilość, tym wyższe prawdopodobieństwo, że istnieje poważna przeszkoda dla przepływu płynu;
- „nasogastric reflux PH” – wartość numeryczna, określa PH treści żołądkowej, skala od 0 do 14, norma mieści się między 3 a 4;
- „rectal examination – feces” – ilość kału, możliwe wartości: „normal”, „increased”, „decreased”, „absent”, brak może wskazywać na niedrożność jelit;
- „abdomen” – określa stan jelit, możliwe wartości: „normal”, „other”, „firm feces in the large intestine”, „distended small intestine”, „distended large intestine”;
- „packed cell volume” – liczba czerwonych krwinek, atrybut numeryczny, norma powinna wynosić między 30 a 50, podwyższony poziom może mieć związek z gorszym krążeniem lub odwodnieniem;
- „total protein” – ilość białka całkowitego, wartość numeryczna, norma wynosi między 6 a 7,5 (gms/dL), podwyższona wartość sugeruje odwodnienie;
- „abdominocentesis appearance” – wygląd płynu pobranego z jamy brzusznej, możliwe wartości: „clear”, „cloudy”, „serosanguinous”; prawidłowy powinien być klarowny, inny może wskazywać uszkodzenie jelit;
- „abdomcentesis total protein” – ilość białka całkowitego z pobranego płynu z jamy brzusznej, wartość numeryczna, wartości w podane w gms/dL; im wyższy poziom, tym większe prawdopodobieństwo, że jelita są uszkodzone;
- „surgical lesion” – czy problem był możliwy do leczenia operacyjnego, stwierdzone albo w momencie podjęcia operacji, albo przy sekcji zwłok, możliwe wartości: „yes”, „no”;
- „lesion 1”, „lesion 2”, „lesion 3” - typy problemu / uszkodzeń; atrybut złożony z czterech liczb:
  - pierwsza określa miejsce uszkodzenia (0 – brak, 1 – żołądek, 2 – jelito cienkie, 3 – jelito grube, 4 – jelito grube i ślepe, 5 – jelito ślepe, 6 – okrężnica poprzeczna, 7 – okrężnica zstępująca, 8 – macica, 9 – pęcherz moczowy, 11 – wszystkie miejsca w jelitach),
  - druga to typ (1 – nieskomplikowany, 2 – zduszenie, 3 – stan zapalny, 4 – inny),
  - trzecia to podtyp (0 – nie dotyczy, 1 – mechaniczny, 2 – porażony/sparaliżowany),
  - czwarta to szczegółowy kod (0 – nie dotyczy, 1 – zatkanie, 2 – wrodzony, 3 – na skutek czynników zewnętrznych, 4 – osłabienie, 5 – skręt, 6 – wgłobienie (część jelita wsuwa się w sąsiednią część jelita), 7 – problem zakrzepowo – zatorowy, 8 – przepuklina, 9 – tłuszczak / uwięźnięcie śledziony, 10 – przemieszczenie)
- „cp data” - czy dla tego przypadku dostępne są dane patologiczne, możliwe wartości: „yes”, „no”.

Przykłady ze zbioru, interpretacja wartości

Ze zbioru uczącego wybrano trzy obserwacje (dziesiątą, trzydziestą i trzydziestą piątą). Ze względu na dużą ilość atrybutów, dane przedstawiono w kilku tabelach: 1, 2, 3, 4. Każda obserwacja posiada inną wartość atrybutu decyzyjnego.
Porównując ze sobą przykłady dla atrybutów zestawionych w Tabeli 1, wszystkie przykładowe obserwacje dotyczą dorosłych zwierząt, z czego jedno z nich było operowane. Koń, który przeżył miał podwyższoną temperaturę. Żaden osobnik w momencie badania nie miał pulsu mieszczącego się w normie, jednak dla zwierzęcia, które ostatecznie przeżyło, było ono widocznie niższe, podobnie jak częstość oddechów. Wszystkie konie miały chłodne lub zimne kończyny oraz obniżony puls obwodowy.

nr	surgery	age	hospital number	rectal temp	pulse	respiratory rate	temp. of extremities	peripheral pulse	outcome
10	yes	adult	528548	38.1	66.0	12.0	cool	reduced	lived
30	no	adult	529475	37.7	96.0	30.0	cool	reduced	died
35	no	adult	528812	NA	104.0	24.0	cold	reduced	euthanized

Tabela 1: Przykłady ze zbioru, porównanie na atrybutach: „surgery", „age", „hospital number", „rectal temperature", „pulse", „respiratory rate", „temperature of extremities", „peripheral pulse"

nr	capillary refill time	pain	peristalsis	abdominal distention	nasogastric tube	nasogastric reflux	nasogastric reflux ph	outcome
10	< 3 sec	mild	hypomotile	none	slight	none	3.0	lived
30	> 3 sec	extreme	absent	severe	significant	< 1 liter	4.0	died
35	> 3 sec	severe	absent	moderate	NA	> 1 liter	NA	euthanized

Tabela 2: Przykłady ze zbioru, porównanie na atrybutach: „capillary refill time", „pain", „peristalsis", „abdominal distention", „nasogastric tube", „nasogastric reflux", „nasogastric reflux ph"

Dla danych przedstawionych w Tabeli 2, dla koni, które ostatecznie nie przeżyły, zanotowano wyższy czas napełniania się naczyń włosowatych, zdawały się również odczuwać większy ból, a ich perystaltyka jelit została całkowicie zahamowana, podczas gdy dla zwierzęcia, które przeżyło, była ona obniżona, ale wciąż obecna. Dla koni, u których perystaltyka jelit była nieobecna, zaobserwowano również wzdęcie brzucha oraz obecność refluksu. U dwóch osobników, dla których zbadano ph treści żołądkowej, nie wystąpiły odchylenia od normy.

nr	mucous membrane	rectal exam (feces)	abdomen	packed cell volume	total protein	abdomo appearance	abdomo protein	outcome
10	bright red	increased	distend large	44.0	6.0	cloudy	3.6	lived
30	pale cyanotic	absent	distend large	66.0	7.5	NA	NA	died
35	pale pink	NA	other	73.0	8.4	NA	NA	euthanized

Tabela 3: Przykłady ze zbioru, porównanie na atrybutach: „mucous membrane", „rectal examination – feces", „abdomen", „packed cell volume", „total protein", „abdominocentesis appearance", „abdomcentesis total protein"

Porównując ze sobą przykładowe obserwacje na atrybutach zestawionych w Tabeli 3, żadne zwierzę nie miało prawidłowego koloru śluzówek. Koń, który przeżył, charakteryzował się niższą liczbą czerwonych krwinek na objętość krwi, pozostałe dwa mają tę wartość podwyższoną. Zwierzę, które ostatecznie zostało uśpione, miało też większą ilość białka całkowitego, czego nie zaobserwowano dla pozostałych dwóch przypadków – obie wartości mieszczą się w normie. Dla jednej obserwacji wykonano badanie polegające na pobraniu płynu z jamy brzusznej, który wykazał nieprawidłowy, mętny kolor, było obecne w nim również białko.

nr	surgery	sugical lesion	lesion 1	lesion 2	lesion 3	cp data	outcome
10	yes	yes	2124	0	0	yes	lived
30	no	yes	4205	0	0	no	died
35	no	yes	7111	0	0	no	euthanized

Tabela 4: Przykłady ze zbioru, porównanie na atrybutach dotyczących występujących uszkodzeń, możliwej i wykonanej operacji oraz obecności danych patologicznych

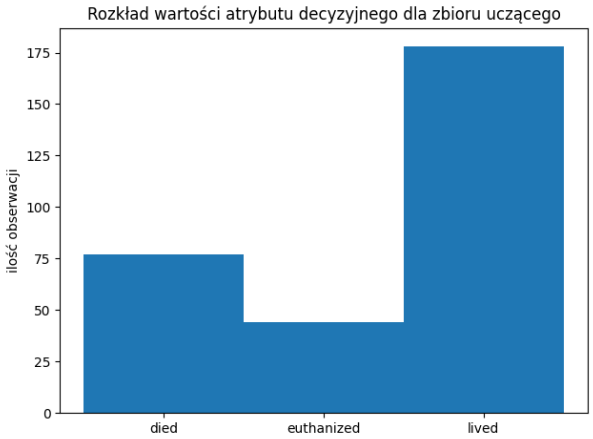
Dla wszystkich przedstawionych obserwacji dla danych zestawionych w Tabeli 4 wystąpiło po jednym uszkodzeniu, które można było operować (z czego podjęto się takiego działania tylko w jednym przypadku). Koń, który przeżył, miał osłabione (porażone) jelito cienkie, problem był jednak nieskomplikowany.

Osobnik, który zmarł, miał zduszone jelito grube oraz ślepe spowodowane przez skręt. Zwierzę uśpione miało zatkaną mechanicznie okrężnicę zstępującą.

2 Wstępna eksploracja zbioru

2.1 Zbalansowanie danych

Badając rozkład wartości na atrybucie decyzyjnym (Rys. 1) okazało się, że zbiór uczący nie jest zbalansowany i zawiera zdecydowaną przewagę obserwacji, dla których zwierzę przeżyło chorobę. Najmniej liczna jest klasa, gdzie koń zostaje uśpiony.

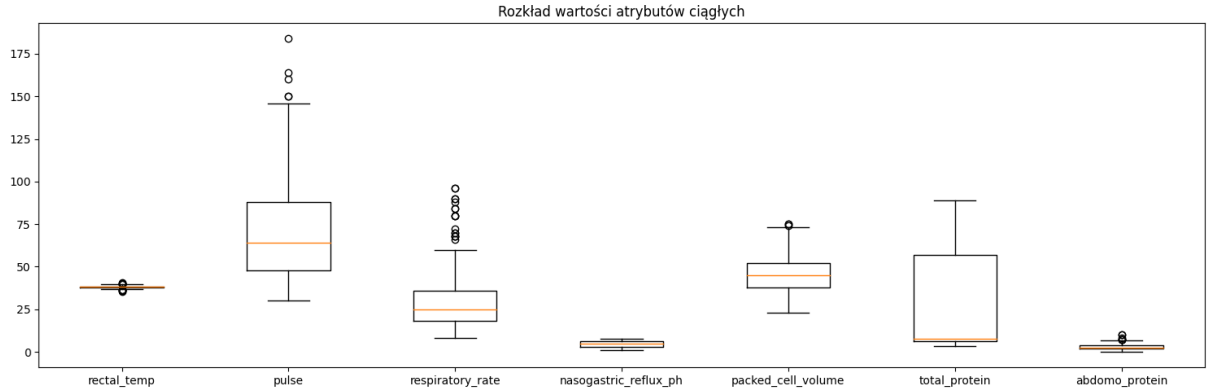


Rysunek 1: Rozkład wartości atrybutu decyzyjnego na zbiorze uczącym

2.2 Rozkład wartości atrybutów

Atrybuty o charakterze ciągłym

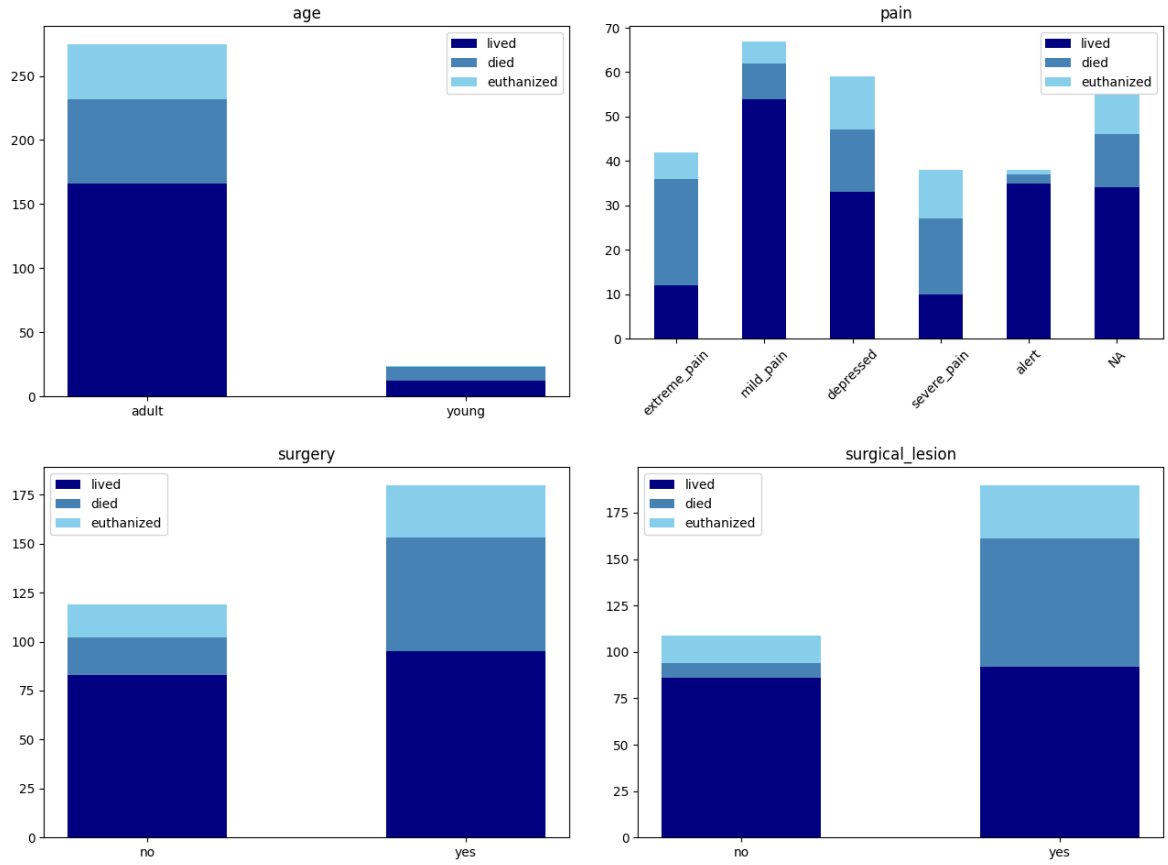
Zakresy atrybutów numerycznych (Rys. 2) są zróżnicowane. Największa rozpiętość wartości występuje na atrybucie „pulse", a najmniejsza na „rectal_temp" (różnica między maksymalną a minimalną wartością wynosi 5,4). Dane na omawianych atrybutach nie mają wielu wartości odstających – najwięcej (17) jest ich na atrybucie „respiratory rate".



Rysunek 2: Zakresy i rozkłady wartości atrybutów ciągłych

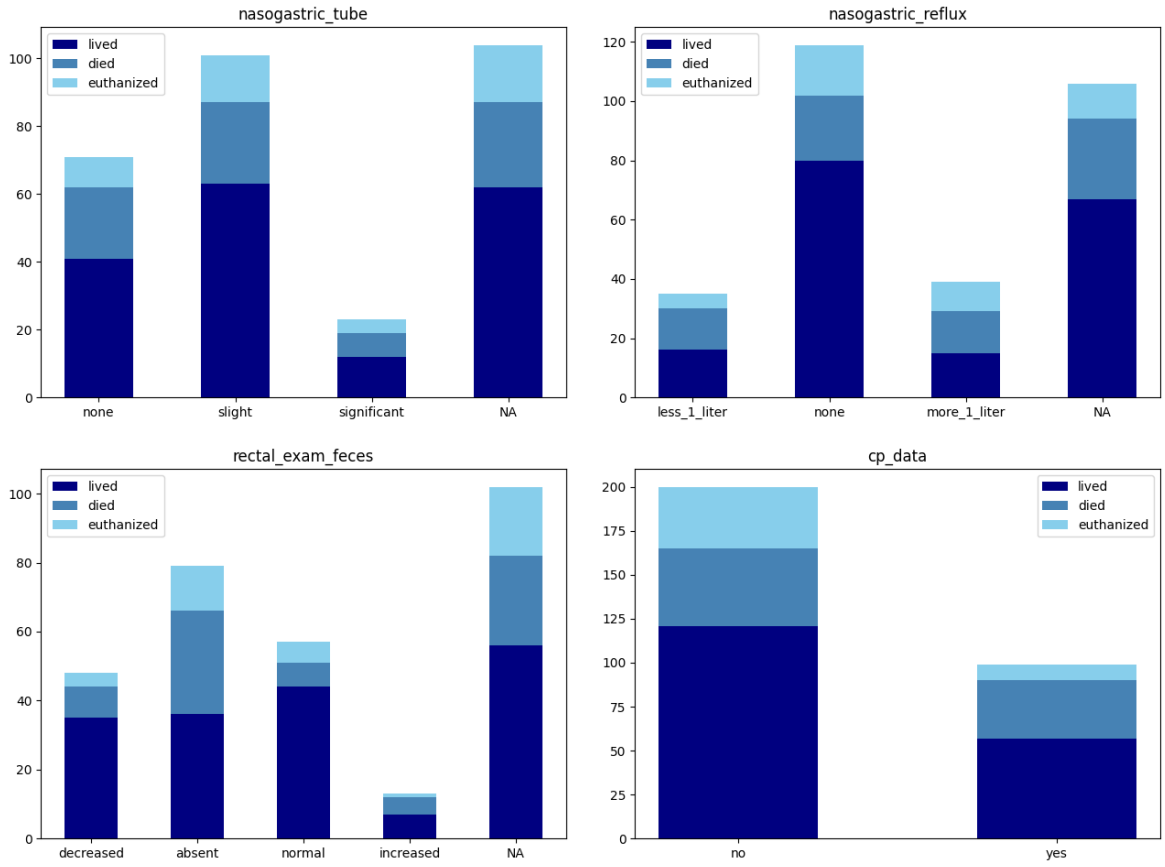
Atrybuty kategoryczne

Większość przypadków w zbiorze dotyczy koni co najmniej półrocznych. Dla atrybutu dotyczącego bólu (Rys. 3) można zauważyć, że zwierzęta, które zdawały się go nie odczuwać lub występował on w małym stopniu, większość osobników przeżywała, a w przypadku, gdy był on poważny (wartości „extreme" oraz „severe") tendencja jest odwrotna. Mimo, że wartości atrybutu są miarą subiektywną, a sam ból nie jest mierzalny, to w naturze tego gatunku leży ukrywanie złego samopoczucia (aby nie być postrzeganym przez drapieżnika jako łatwym do zaatakowania). Jeżeli więc widoczne jest cierpienie, to odczuwany ból jest na tyle duży, że zwierzę nie jest w stanie go zamaskować, co świadczy o złym stanie. Większość przypadków choroby mogła być leczona operacyjnie, u więcej niż połowy przeprowadzono takie działanie.



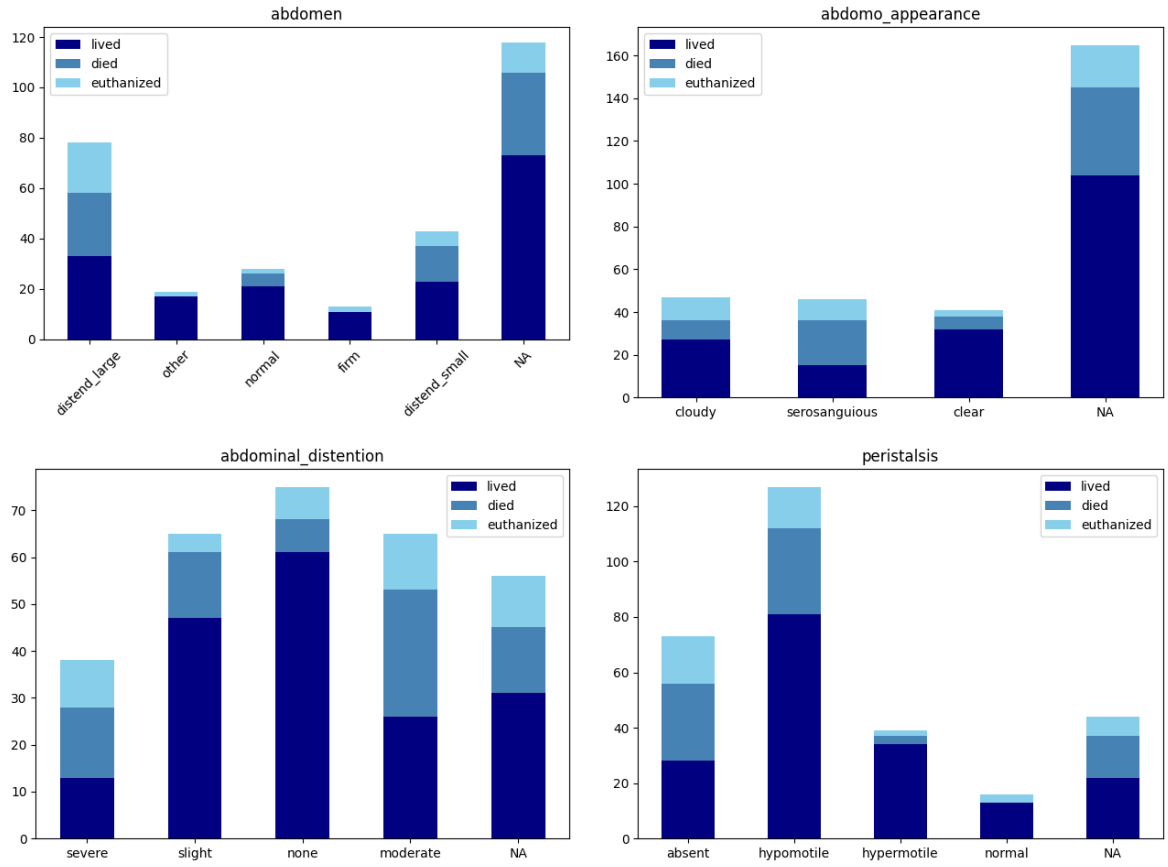
Rysunek 3: Rozkłady wartości na atrybutach kategorycznych: „age", „pain", „sugery", „surgical lesion"

Na atrybutach „nastogastric tube”, „nasograstic reflux” oraz „rectal examination – feces” występuje wiele braków w danych (Rys. 4). Konie, u których występował refluks, przeżywały rzadziej, niż gdy takowy nie występował. Taka zależność występuje również w przypadku, gdy zatrzymana jest produkcja kału – co może świadczyć o zahamowaniu pracy układu pokarmowego lub zaplątaniu/zapchaniu jelita, które jest niebezpieczne dla życia, gdyż może prowadzić nawet do pęknięcia narządu i może nie być wyleczalne bez operacji.



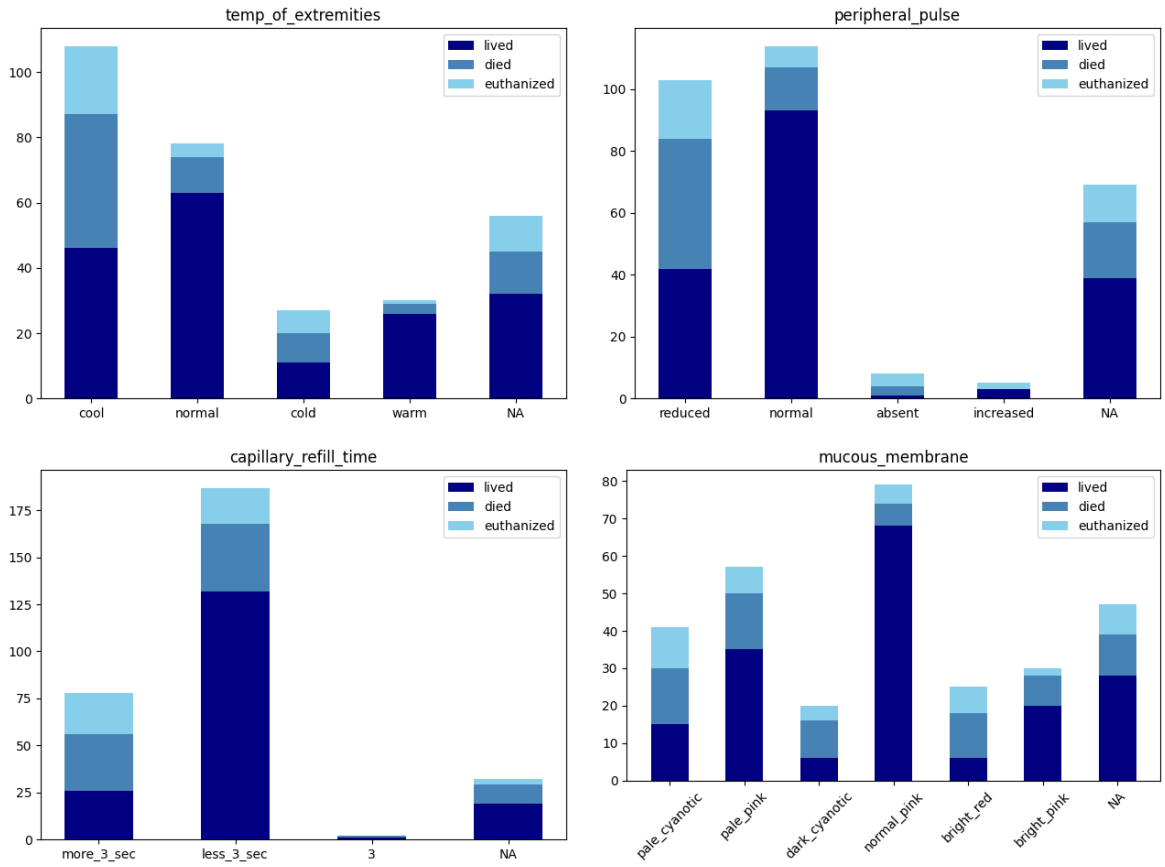
Rysunek 4: Rozkłady wartości na atrybutach kategorycznych: „nastogastric tube”, „nasograstic reflux”, „rectal examination – feces”, „cp data”

Większości koni nie robiono badania polegającego na pobraniu i zbadaniu płynu z jamy brzusznej, lub wynik takowego nie był notowany (Rys. 5). Dla obserwacji, które zawierały dane na ten temat można zauważyć, że w przypadku płynu surowiczego występuje większy odsetek zwierząt, które zmarły, a dla klarownego większość przeżywa. Mniej koni przezwyciężało chorobę, jeżeli występowały wzdęcia brzucha, w porównaniu z tymi, u których nie występowały. Większość zwierząt miało spowolnioną lub nieobecną perystaltykę jelit.



Rysunek 5: Rozkłady wartości na atrybutach kategorycznych dotyczących brzucha

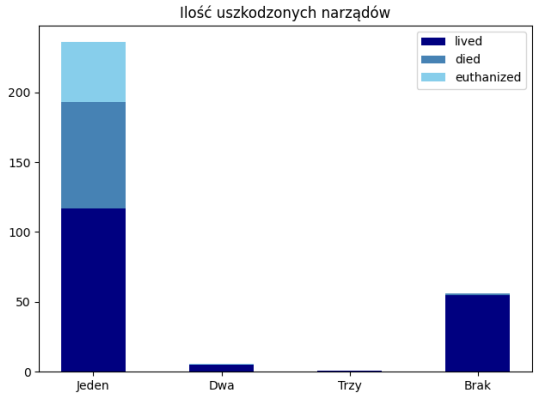
Zwierzęta, u których zaobserwowano cechy wskazujące na gorsze/niewydolne krążenie (Rys. 6), rzadziej przezwyciężały chorobę – dotyczy to zarówno przypadków, gdzie czas napełniania się naczyń włosowatych był dłuższy, jak i temperatury kończyn (chłodne i zimne), a także pulsu obwodowego (dla obniżonego występuje więcej obserwacji, gdzie zwierzę nie przeżywa, a w przypadku, gdy jest ono niewyczuwalne, większość nie przeżywa). Nieprawidłowy kolor śluzówek (siny lub czerwony) również zawiera mniej przypadków, gdy koń przeżywa.



Rysunek 6: Rozkłady wartości na atrybutach kategorycznych dotyczących krążenia

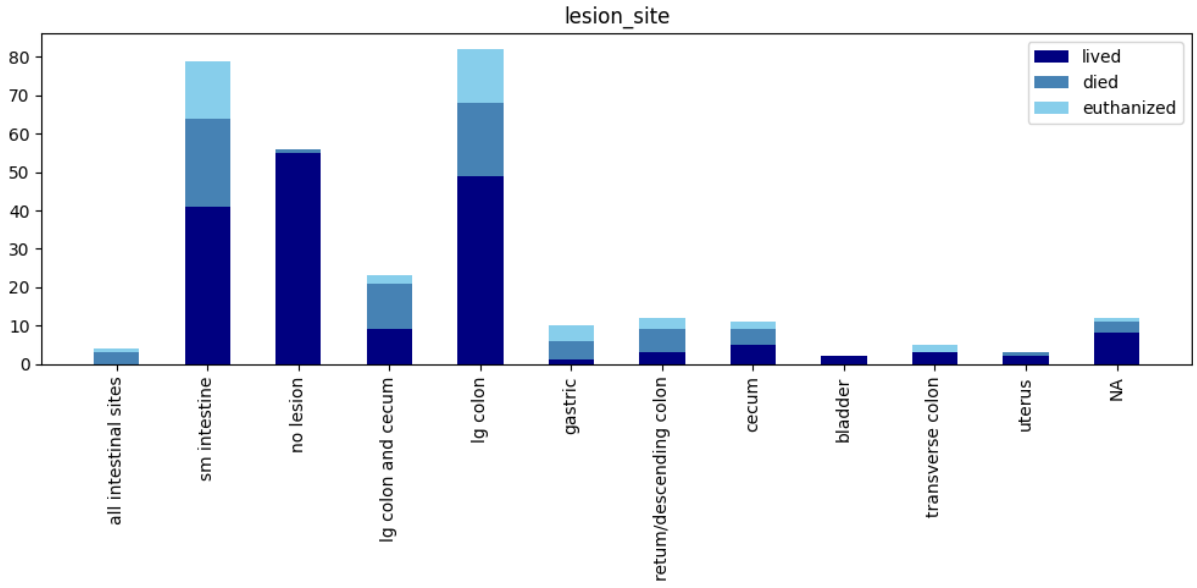
Atrybuty opisujące uszkodzenia („lesion 1”, „lesion 2”, „lesion 3”)

Na trzech atrybutach opisujących uszkodzenia (schorzenia) narządów, u większości koni występowało tylko jedno. Dwa występują w sześciu obserwacjach, a trzy tylko w jednej (Rys. 7).



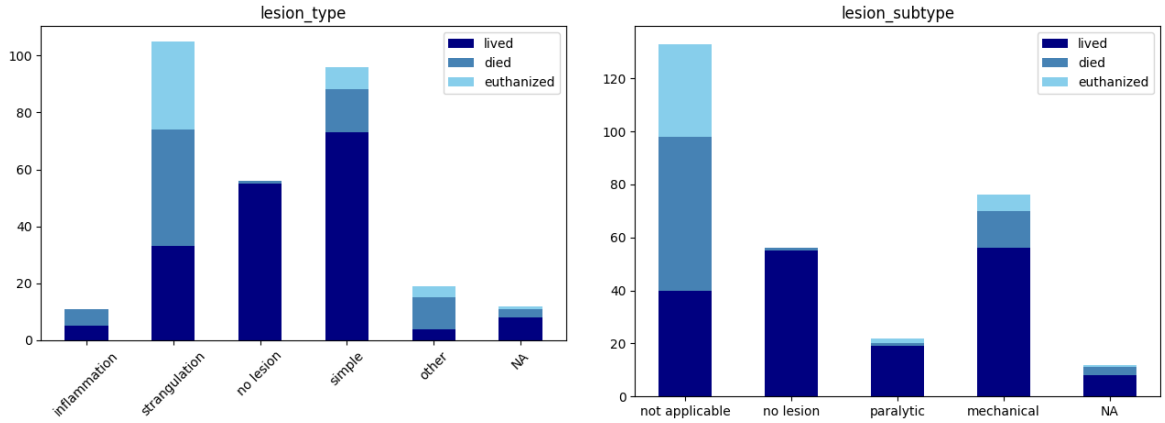
Rysunek 7: Ilość chorych (uszkodzonych) narządów w zbiorze treningowym

Przeanalizowana została kolumna z wartościami atrybutu „lesion 1", w celu dokładniejszego zbadania występujących schorzeń. Przekształcono ją do postaci czterech atrybutów; „lesion site" – umiejscowienie uszkodzenia (Rys. 8), „lesion code" – przyczyna problemu (Rys. 10), „lesion type" oraz „lesion subtype" – określające typy oraz skomplikowanie problemu (Rys. 9). W przypadku, gdy liczba była zapisana nieprawidłowo, na wszystkich kolumnach wpisywano brak danych.

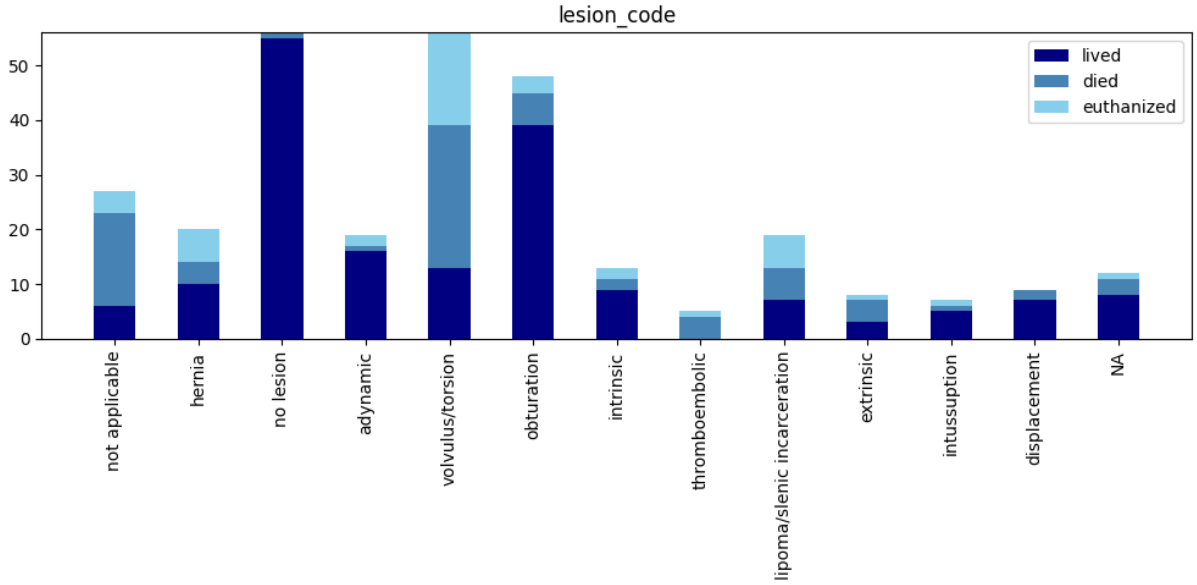


Rysunek 8: Rozkład wartości umiejscowienia schorzenia

W przypadku braku zmian chorobowych, zwierzę prawie zawsze przeżywało. Najczęściej dochodziło do uszkodzenia jelita grubego lub cienkiego, przy czym przeżywalność w obu przypadkach jest podobna. Jeżeli poza jelitem grubym, chora była również kątnica, to prawdopodobieństwo na przeżycie spadało (aczkolwiek dla tego przypadku występuje mało obserwacji). Najrzadsze były schorzenia pęcherza moczowego oraz macicy (które w większości zakończyły się wyleczeniem), a także uszkodzenia rozległe w jelitach, które zawsze kończyły się śmiercią.



Rysunek 9: Rozkład wartości typów schorzeń

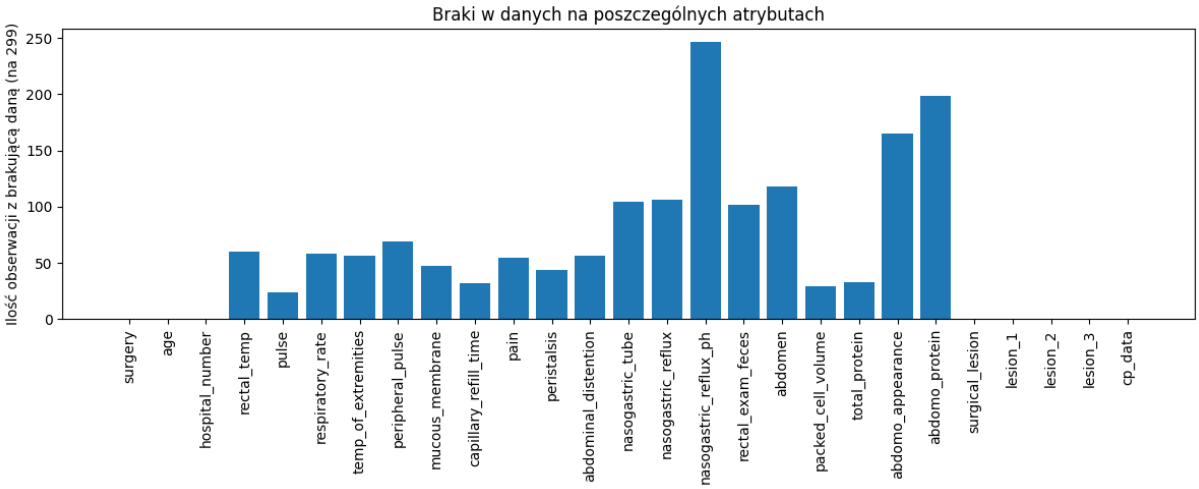


Rysunek 10: Rozkład wartości przyczyn schorzeń

Śmiertelne w dużej mierze były również schorzenia, przy których uwięziona była śledziona lub występował tłuszczak, a także występował problem zakrzepowo-zatorowy (Rys. 10). Stosunkowo łatwe do wyleczenia, a zarazem jednymi z częstszych wydają się schorzenia, których przyczyną jest niedrożność. Jednym z najczęstszych oraz najbardziej śmiertelnych przyczyn jest skręt – występuje w tym przypadku nie tylko duży odsetek śmierci, ale też eutanacji.

2.3 Braki w danych

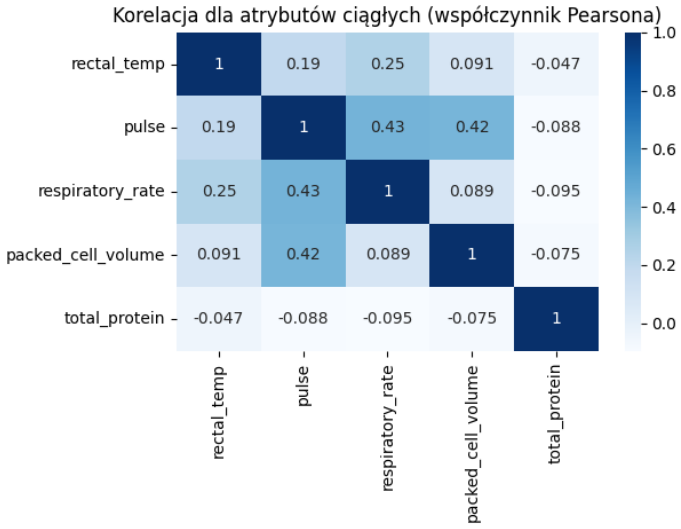
Większość atrybutów w zbiorze posiada brakujące dane, przy czym najrzadziej wykonywano badanie PH treści żołądkowej (ilość brakujących danych to 246). Stosunkowo nieczęsto pobierano płyn z jamy brzusznej w celu określenia jego koloru (brak 165 danych), a także badano w nim poziom białka (brak 198 danych).



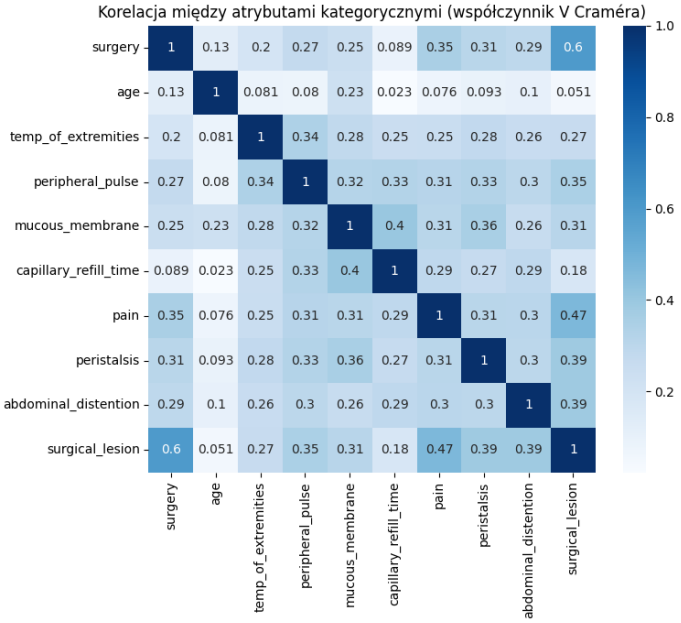
Rysunek 11: Ilość brakujących danych dla każdego atrybutu

2.4 Zależności między atrybutami

Zmierzono zależność między atrybutami ciągłymi (używając współczynnika Pearsona) oraz kategorycznymi (stosując współczynnik V Craméra). Pomięte zostały atrybuty, dla których wiele (co najmniej 80) wartości było brakujących.
Wśród atrybutów numerycznych, największą korelację zanotowano pulsu oraz częstości oddechów oraz między pulsem a ilością czerwonych krwinek na jednostkę objętości (Rys. 12). Najmniej skorelowany z innymi cechami był atrybut opisujący białko całkowite.



Rysunek 12: Wartości współczynnika Pearsona dla atrybutów ciągłych



Rysunek 13: Wartości współczynnika V Craméra dla atrybutów kategorycznych

Wśród atrybutów kategorycznych, najbardziej skorelowaną parą atrybutów jest możliwe podjęcie leczenia operacyjnie oraz podjęta operacja, gdzie wartość współczynnika wynosi 0,6 (Rys. 13). Możliwe jest, że właściciele koni podejmowali decyzje o operacji na podstawie przesłanek dostarczanych przez weterynarzy (którzy w zależności od nabytego doświadczenia, wstępnie są w stanie stwierdzić, co może być przyczyną problemu). Najmniej skorelowanym z innymi atrybutem był wiek.

3 Wykonane przekształcenia zbioru

Ze względu na dużą ilość atrybutów kategorycznych oraz braków w danych, zbiór uległ następującym przekształceniom (przekształcenia wykonano według zaprezentowanej kolejności):

- usunięto przypadki, w których występował brak danych na co najmniej 13 atrybutach (24 obserwacje),
- usunięto atrybuty, dla których brak danych obejmował co najmniej 80 obserwacji (7 atrybutów)
- usunięto atrybuty, które wydawały się dostarczać małą wartość do problemu: „hospital number", „lesion 2", „lesion 3", a także „cp data", jako że nie dotyczy ona bezpośrednio metryki dotyczącej zdrowia zwierzęcia,
- wartościowym atrybutem wydaje się być „lesion 1", ponieważ zawiera wiele informacji, jednak schorzenie mogło mieć cechy nieokreślone (np. przyczynę problemu, określoną przez czwartą liczbę) – takich przypadków nie można uzupełnić sztucznie danymi, więc w kolumnie pozostawałyby wartości typu „nie dotyczy", problematyczne przy późniejszej klasyfikacji; występowały również nieprawidłowo zapisane dane, dla których np. jedna z liczb była brakująca; ostatecznie kolumnę również usunięto, ze względu na trudności w jej przekształceniu.
- atrybuty kategoryczne, których wartości nie da się przedstawić na skali porządkowej („surgery", „pain", „surgical lesion") zostały przekształcone za pomocą OneHotEncoder; ewentualne braki w danych na atrybucie „pain" skutkowały wpisaniem wartości 0 na wszystkich kolumnach, jakie powstały przy przekształcaniu tego atrybutu;
- atrybuty kategoryczne, których wartości można przedstawić na skali porządkowej zostały przekształcone za pomocą OrdinalEncoder, przy czym dla poszczególnych atrybutów, uporządkowanie wartości wyglądało następująco:
  - „age": „young", „adult"
  - „temp of extremities": „cold", „cool", „normal", „warm"
  - „peripheral pulse": „absent", „reduced", „normal", „increased"
  - „mucous membrane": „dark cyanotic", „pale cyanotic", „pale pink", „normal pink", „bright pink", „bright red"
  - „capillary refill time": „less 3 sec", „3", „more 3 sec"
  - „peristalsis": „absent", „hypomotile", „normal", „hypermotile"
  - „abdominal distention": „none", „slight", „moderate", „severe"
- atrybuty kategoryczne oraz ciągłe, dla których występowały braki w danych, zostały uzupełnione za pomocą KNNImputer z atrybutem *weights="distance"*

4 Testy

Do analizy wybrano następujące algorytmy z biblioteki *scikit–learn*:

- `RandomForestClassifier`, ze względu na dużą ilość danych kategorycznych struktura drzewa wydaje się być użyteczna;
- `SVC`, ze względu na to, że dobrze potrafi sobie radzić z niezbalansowanymi i wielowymiarowymi danymi;
- `VotingClassifier` z estymatorami: drzewem decyzyjnym (`DecisionTreeClassifier`), klasyfikatorem k–najbliższych sąsiadów (`KNeighborsClassifier`) oraz opartym o sieci neuronowe (`MLPClassifier`); drzewo decyzyjne efektywnie potrafi wyodrębnić istotne atrybuty oraz wykryć nieliniowe zależności, podczas gdy sieć neuronowa potrafi odnaleźć skomplikowane, globalne zależności, a klasyfikator k–najbliższych sąsiadów działa w bardziej „lokalny" sposób.

W celu wybrania najlepszej konfiguracji dla każdej metody wykorzystano `GridSearchCV`. Poza trafnością klasyfikacji, notowano również wartości metryk przydatnych do oceny modelu przy niezbalansowaniu danych: g–mean oraz F1–score (liczona jest średnia ważona, `average='weighted'`), przy czym ocena jakości (parametr `refit`) była określana przez F1–score. Przy modelach wrażliwych na brak znormalizowania danych: `SVC` oraz `VotingClassifier` (ze względu na `KNeighborsClassifier`) użyto przeskalowania atrybutów ciągłych za pomocą `StandardScaler`.
Dla `RandomForestClassifier`, testowano następujące parametry:

- maksymalna głębokość (`max_depth`): nieograniczona, 1, 2, 5;
- wagi (`class_weight`): brak, zbalansowane;
- minimalna ilość przypadków w liściu (`min_samples_leaf`): 1, 2, 5;
- minimalna ilość przypadków przy podziale (`min_samples_split`): 2, 5, 10.

Dla `SVC`, testowano następujące parametry:

- rdzeń (`kernel`): rbf, liniowy, wielomianowy;
- stopień wielomianu (`degree`): 1, 3, 5;
- wagi (`class_weight`): brak, zbalansowane.

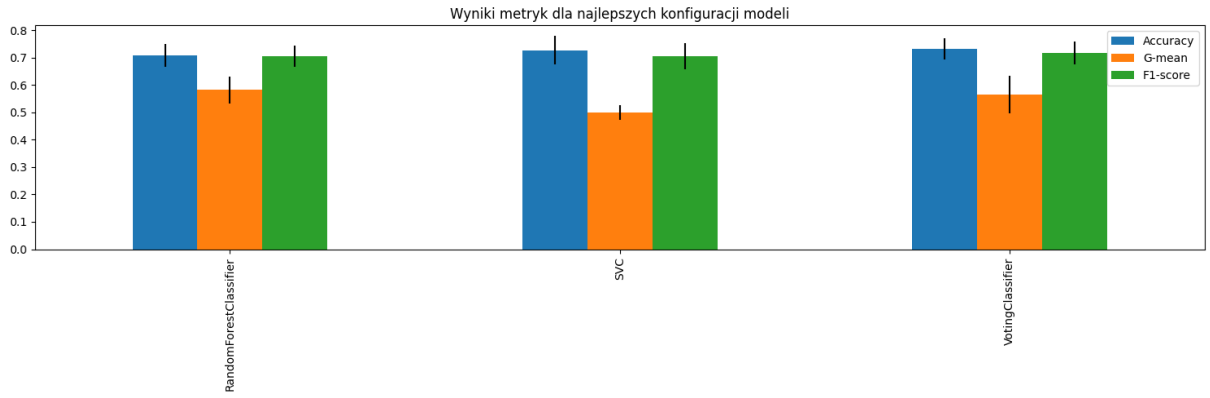
Dla `VotingClassifier`, testowano następujące parametry:

- rodzaj głosowania (`voting`): większościowy, na podstawie przewidywwanych prawdopodobieństw (soft);
- dla drzewa decyzyjnego (`DecisionTreeClassifier`) przetestowano różne głębokości drzewa (`max_depth`): 1, 2, 5, 10;
- dla k–najbliższych sąsiadów (`KNeighborsClassifier`) przetestowano wagi jednolite oraz bazujące na odległościach (`weights`), a także sprawdzono różną ilość sąsiadów (`n_neighbors`): 3, 5, 10;
- dla sieci neuronowej (`MLPClassifier`) sprawdzono obie możliwe wartości parametru (`shuffle`), a także maksymalną liczbę iteracji (`max_iter`) wynoszącą kolejno: 200, 500, 1000;

5 Wyniki

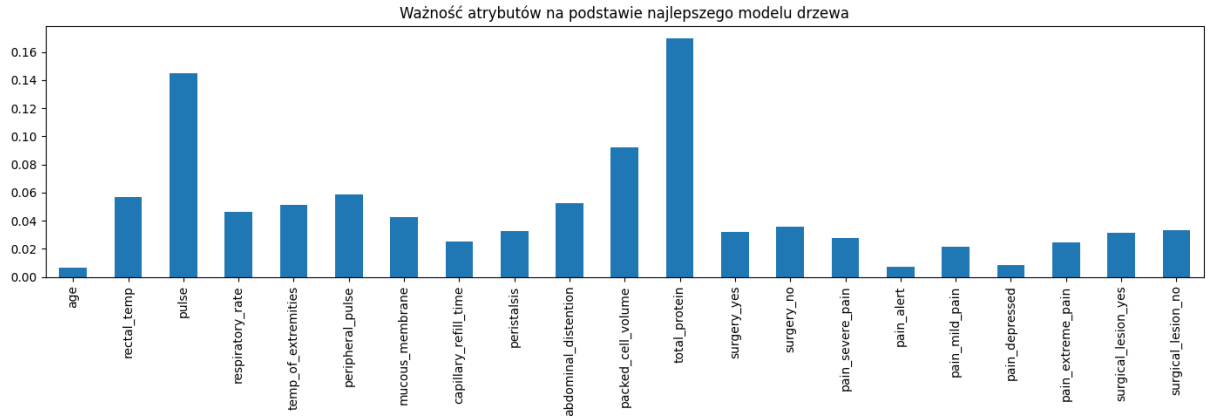
Najlepsze uzyskane konfiguracje miały następujące parametry:

- `SVC`: brak wag, pierwszy stopień wielomianu, liniowy rdzeń;
- `RandomForestClassifier`: zbalansowane wagi, brak ograniczeń głębokości, minimum 10 przypadków do wykonania podziału, minimum 2 obserwacje na liść;
- `VotingClassifier`: głosowanie na podstawie prawdopodobieństw, drzewo decyzyjne o maksymalnej głębokości 2, dla k–najbliższych sąsiadów najlepsze okazało się rozpatrywanie sąsiedztwa 3 obserwacji z wagami bazującymi na odległościach, natomiast sieć neuronowa powinna uczyć się maksymalnie 200 iteracji, a przypadki nie powinny być mieszane.



Rysunek 14: Wyniki metryk uzyskane dla najlepszych modeli

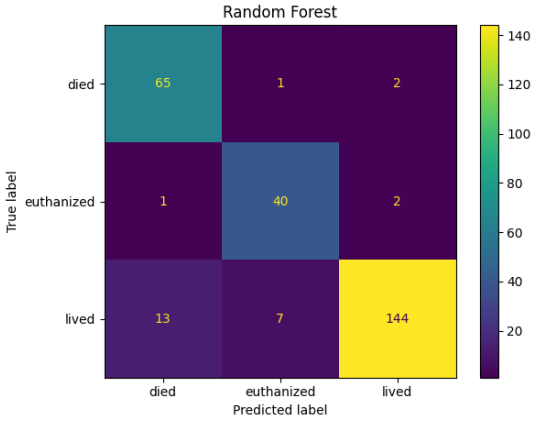
Wartości g–mean sugerują, że model może mieć problemy ze zbytnią generalizacją (Rys. 14), lub niektóre obserwacje są do siebie bardzo zbliżone pod względem wartości atrybutów, więc występuje problem z zaklasyfikowaniem ich do prawidłowej klasy. F1–score ma zbliżone wartości dla wszystkich modeli, wynosi około 0,7. Problem z prawidłowym zaklasyfikowaniem przypadku do odpowiedniej klasy, który może wynikać z wielu przyczyn – sztucznego uzupełniania danych, usuwania kolumn, małej liczby obserwacji lub zbytniego podobieństwa przypadków należących do różnych klas.



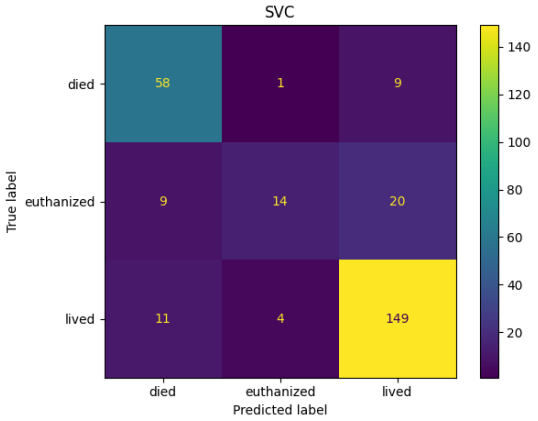
Rysunek 15: Ważność atrybutów dla drzewa decyzyjnego

Drzewo decyzyjne wygenerowane przez `RandomForestClassifier` (Rys. 15) wskazało jako najważniejszy atrybut ilość białka całkowitego („total protein") oraz puls („pulse"). Jedną z ważniejszych cech była też ilość krwinek czerwonych („packed cell volume"). Wartości tych atrybutów nie były ustalane na drodze

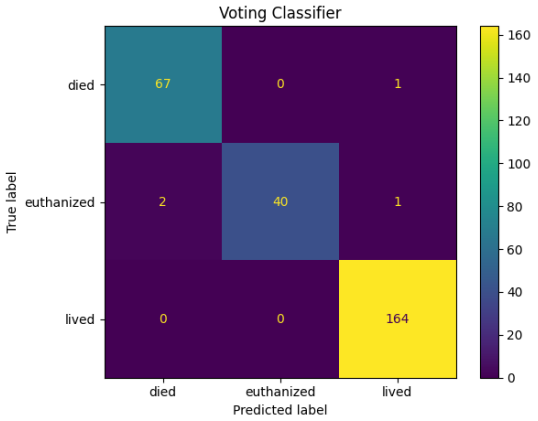
subiektywnej oceny, a są wynikiem przeprowadzonych badań. Braki w danych dla tych cech były jednymi z najniższych, u których takie występowały (Rys. 11). W zestawieniu nie znalazł się żaden atrybut, który byłby uznany za całkowicie nieważny.
Wygenerowano macierze pomyłek dla zbioru uczącego (Rys. 16). Najmniej błędów popełnił `VotingClassifier`, a najwięcej problemów z prawidłowym zaklasyfikowaniem obserwacji ma `SVC`. Drzewo decyzyjne ma tendencję do przewidywania, że zwierzę zmarło, podczas gdy w rzeczywistości przeżyło – możliwe, że istnieją w zbiorze ciężkie przypadki, które ostatecznie pokonały jednak chorobę.



(a) Macierz pomyłek (drzewo decyzyjne)



(b) Macierz pomyłek (SVC)

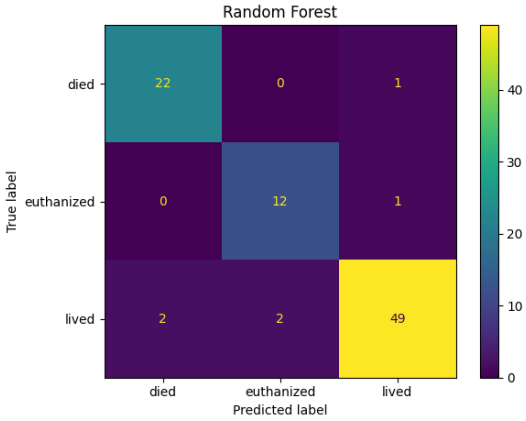


(c) Macierz pomyłek (VotingClassifier)

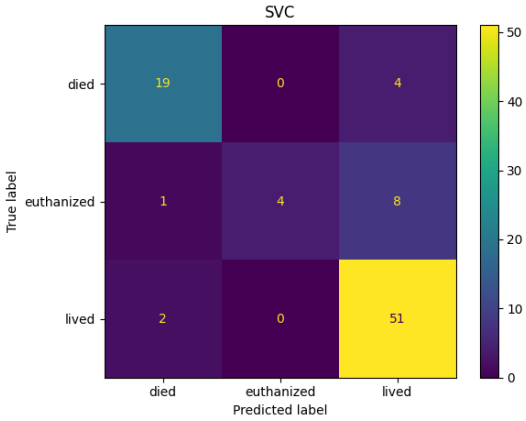
Rysunek 16: Macierze pomyłek dla najlepszych konfiguracji modeli (zbiór uczący)

6 Zbiór testowy - macierze pomyłek

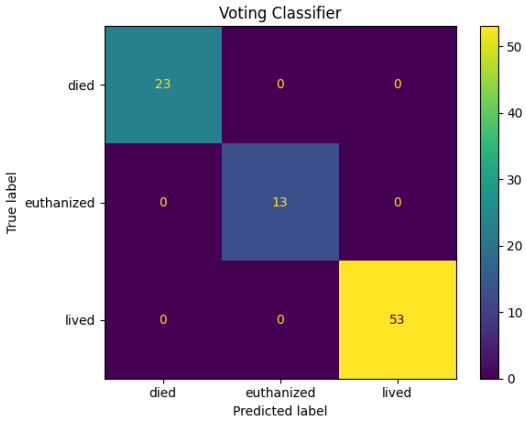
Razem ze zbiorem uczącym, autorzy dostarczyli również zbiór testowy. Wygenerowano dla każdego z trzech modeli macierz pomyłek (Rys. 17). Dla `VotingClassifier` nie wystąpił żaden błąd. Najwięcej błędów popełnił `SVC`, gdzie widoczny jest problem z poprawnym klasyfikowaniem przypadków do klasy „euthanized" – podobnie, jak miało to miejsce przy klasyfikacji wykonanej na zbiorze uczącym (Rys. 16).



(a) Macierz pomyłek (drzewo decyzyjne)



(b) Macierz pomyłek (SVC)



(c) Macierz pomyłek (VotingClassifier)

Rysunek 17: Macierze pomyłek dla najlepszych konfiguracji modeli (zbiór testowy)

7 Wnioski

Wybrany zbiór danych nie był łatwy przez ilość braków i wymagał wiele przekształceń, aby można było z nim pracować. Spodziewanym zjawiskiem była trudność w zaklasyfikowaniu obserwacji w przypadku, gdy koń umierał lub był usypiany, ze względu na to, że decyzję o zakończeniu życia zwierzęcia podejmuje się raczej w momencie, w którym jego stan zdrowia jest raczej zły i bez nadziei na poprawę. Taka sytuacja jednak nie wystąpiła, co sugeruje, że istnieją wyraźne różnice między obserwacjami dla tych klas.
Połączenie ze sobą kilku klasyfikatorów pozwoliło na wykonanie niemal bezbłędnych predykcji, aczkolwiek prosta struktura drzewa również potrafiła efektywnie przewidywać klasę, do jakiej należy dana obserwacja.

Literatura

[1] Bruno Archetti. Horse Colic Dataset. `https://www.kaggle.com/datasets/brunoarchetti/horse-colic-dataset`.

[2] Mogens T. Christophersen, Nana Dupont, Kristina S. Berg-Sørensen, Christel Konnerup, Tina H. Pihl, and Pia H. Andersen. Short-term survival and mortality rates in a retrospective study of colic in 1588 danish horses. *Acta Veterinaria Scandinavica*, 56(1):20, Apr 2014.

[3] Laila Curtis, John H. Burford, Gary C. W. England, and Sarah L. Freeman. Risk factors for acute abdominal pain (colic) in the adult horse: A scoping review of risk factors, and a systematic review of the effect of management-related changes. *PLOS ONE*, 14(7):1–32, 07 2019.

[4] G. A. SUTTON, R. ERTZMAN-GINSBURG, A. STEINMAN, and J. MILGRAM. Initial investigation of mortality rates and prognostic indicators in horses with colic in israel: A retrospective study. *Equine Veterinary Journal*, 41(5):482–486, 2009.