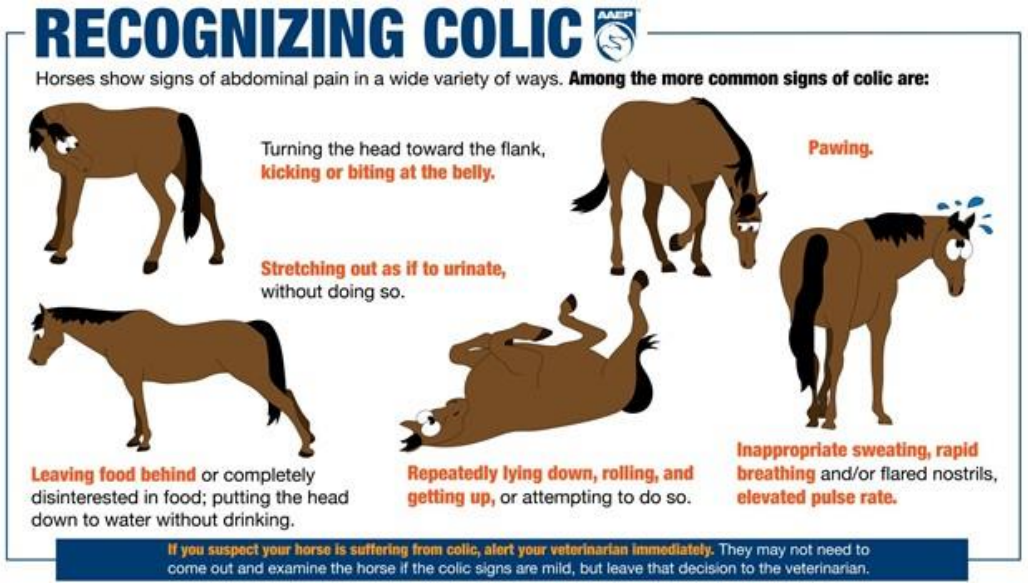
Kolka u konia

Eksploracja i wizualizacja danych

Jakub Popowski

Kolka u konia

Najpoważniejsza choroba konia
- Odpowiada za 32% wszystkich zgonów.
- Co roku na kolkę zachoruję 4.2% koni.



Zbiór danych
- Zbiór danych z repozytorium UCI – Horse Colic Data Set.
- 368 wierszy
- 28 atrybutów
- Około 30% brakujących wartości.
- Zbiór opublikowany w roku 1989.

Cel eksperymentu

Stworzenie modelu, który na podstawie danych wejściowych określi czy koń chory na kolkę:
- przeżyje,
- umrze,
- zostanie poddany eutanazji.

Przygotowanie danych

Wartości brakujące:
- usunięto 8 kolumn i 1 wiersz,
- zastąpione średnią dla cech ciągłych,
- zastąpione modą dla cech dyskretnych,
- wyznaczone z KNN dla jednej kolumny.

Kolumny dotyczące lezji (zmian chorobowych):
- po 3 na konia (większość obserwacji ma 0 lub 1),
- zamieniono na liczbe lezji i ostatnią lezje,
- każda lezja składa się z 4 cech, zakodowanych pozycyjnie jako cyfra(y) w liczbie,
- liczba cyfr na cechę nie jest stała (1 i 4 cecha ma 1 lub 2 cyfry),
- niejednoznaczność – poczyniono pewne założenia.

Wszystkie cechy jakościowe zakodowane jako liczby:
- przywrócony naturalny porządek dla cech porządkowych,
- zdekodowano cechy nominalne do reprezentacji tekstowej,
- cechy nominalne zakodowano jako *dummy variables*.

Utworzono Pipeline:

```
dataTransformationPipe = Pipeline([
    ('column_dropper', columnDropper),
    ('row_dropper', rowDropper),
    ('fill_na', fillNaTransformer),
    ('discrete_encoding', discreteEncodingTransformersPipe),
    ('lesion', lesionPipe),
    ('one_hot', oneHotEncodersTransformer),
    ('scaler', minMaxScalerTransformer)
])
```

Brak wartości odstających dla cech ciągłych.

Dane zostały przeskalowane za pomocą min-max.

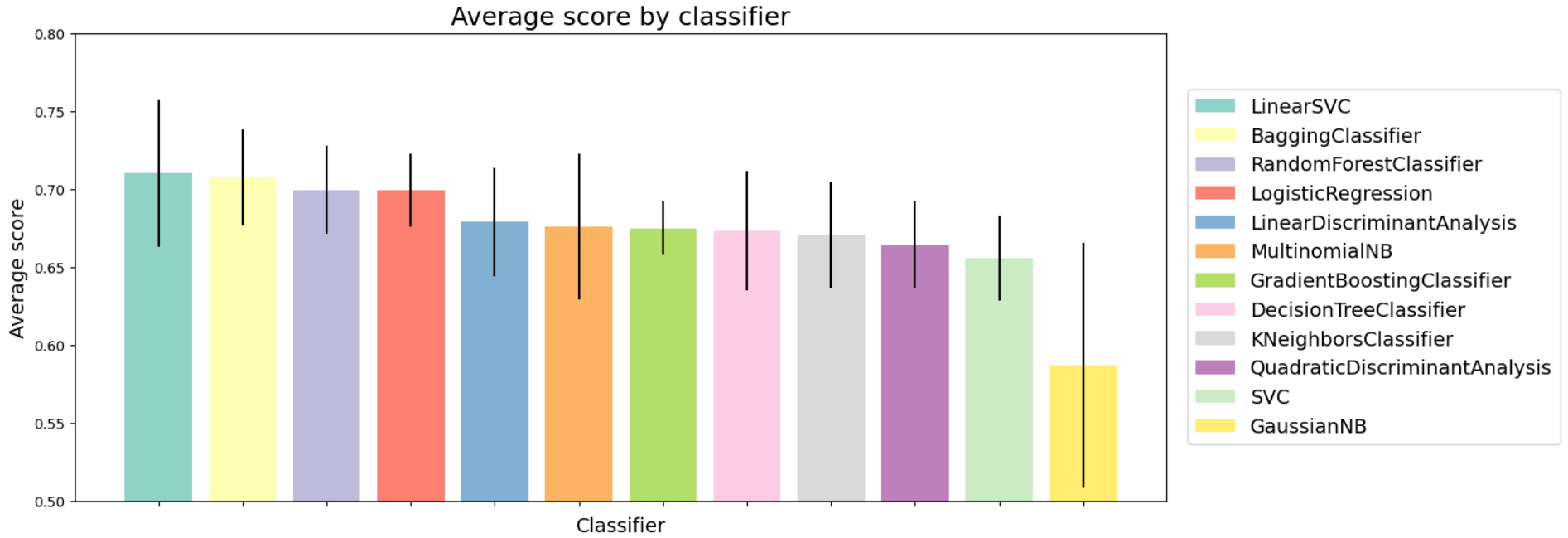
Modelowanie

Wybrano miarę F1 micro jako miarę ewaluacji.

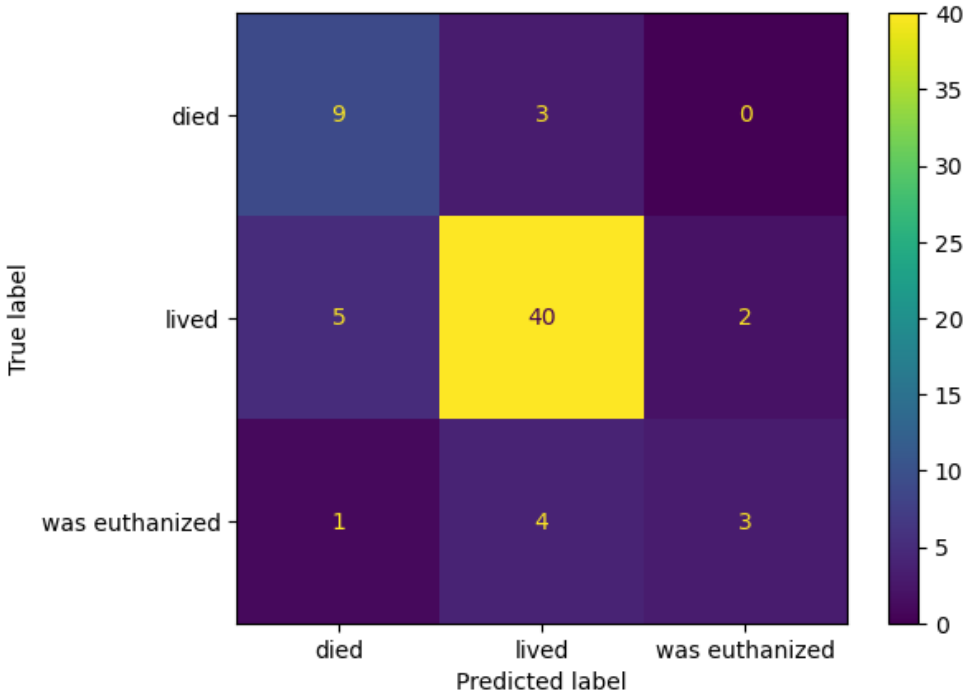
Jako ostateczny model wybrano <u>las losowy</u>, ponieważ spośród 4 zdecydowanie najlepszych klasyfikatorów ma on:

- 3 najlepszy średni wynik,
- najlepszy ‚najlepszy wynik',
- 2 najlepsze średnie STD.

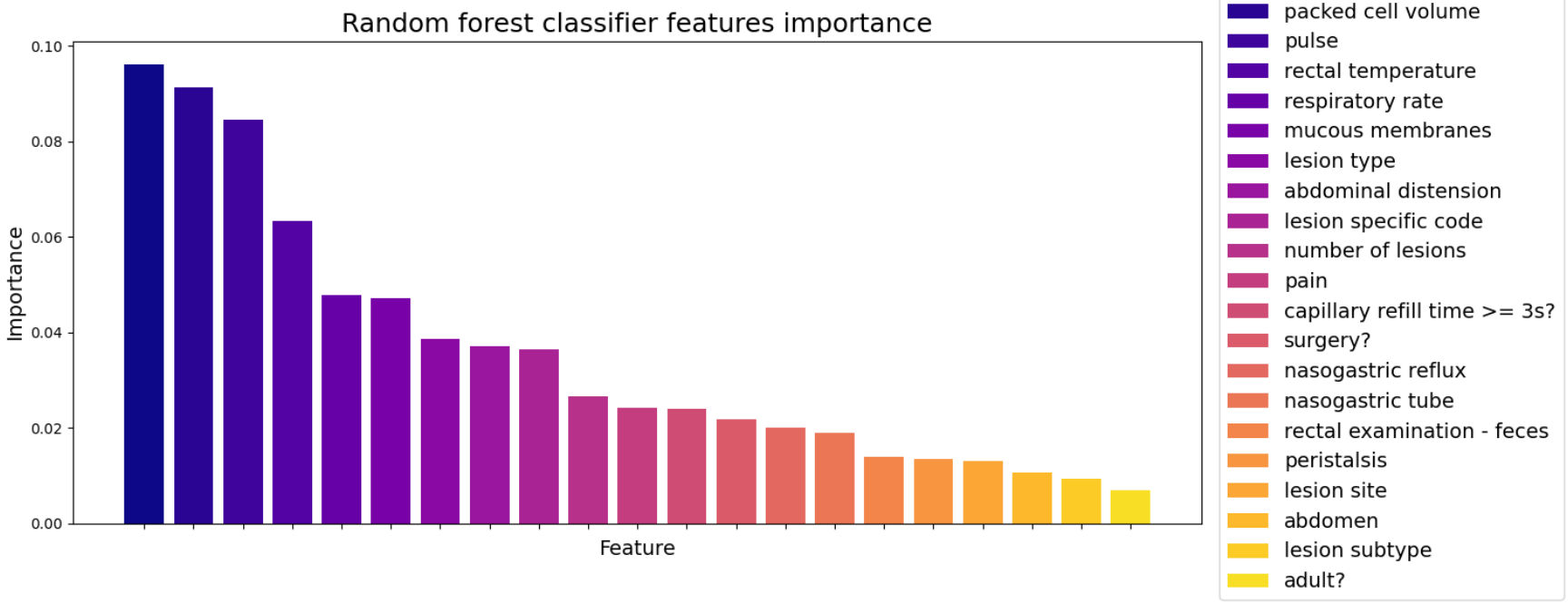
	Klasyfikator	Najlepszy wynik	Średni wynik	Średnie STD
1	LinearSVC	0.7260	0.7099	0.0473
2	BaggingClassifier	0.7224	0.7074	0.0308
3	**RandomForestClassifier**	**0.7324**	**0.6994**	**0.0283**
4	LogisticRegression	0.7025	0.6991	0.0233



Ewaluacja



Wynik F1 micro ewaluacji na zbiorze testowym wynosi <u>0.7761</u>.



Bonus – klasyfikacja binarna

Mimo, że zadanie to klasyfikacja wieloklasowa to zasadniczym pytaniem jest to czy koń przeżyje czy umrze. W związku z tym, jako bonus, utworzono model klasyfikacji binarnej:

- wynik *was euthanized* zamieniono na *died*,
- wytrenowano klasyfikator lasu losowego z takimi samymi hiperparametrami.

Wynik F1 dla klasyfikacji binarnej wynosi <u>0.8060</u>.