Adam Przybyłek, nr indeksu: 105952
Piotr Gawroński, nr indeksu: 81175

WYBRANE METODY UCZENIA MASZYNOWEGO

PORÓWNANIE METOD KLASYFIKACJI - APLIKACJA SHINY

SPIS TREŚCI:

1. OPIS PROJEKTU

Projekt to aplikacja shiny, która porównuje wybrane algorytmy klasyfikacji (knn, ctree oraz naiveBayes) dla różnych zbiorów danych (IRIS, INDIAN, ABALONE)

knn - jest to algorytm k najbliższych sąsiadów. Pochodzi z pakietu *class*, dodatkowo pobiera parametr *k* (jest to liczba sąsiadów), który zawiera się w przedziale od 1 do 10.

ctree - jest to algorytm drzewa decyzyjnego, pochodzi z pakietu *party*, dodatkowo przyjmuje parametr *minsplit*, który jest obliczany dla każdego zestawu danych oddzielnie i przyjmuje wartości od 1 do 1/2 liczności zbioru trenującego.

naiveBayes - jest to algorytm "naiwnego Bayesa", pochodzi z pakietu *e1071*, dodatkowo przyjmuje parametr *laplace* (tzw. poprawka laplaca), który zawiera się w przedziale od 0 do 100.

IRIS - zbiór danych opisujący irysy. Pochodzi z pakietu datasets. Zawiera 150 obiektów oraz 5 atrybutów. Zmienną decyzyjną jest *Species* (gatunek irysa) i przyjmuje ona 3 klasy: *setosa*, *virginica* oraz *versicolor*

INDIAN - zbiór danych opisujący ludzi w Indiach chorych na cukrzycę. Pobrany został ze strony: https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data
Zawiera 768 obiektów oraz 9 atrybutów. Zmienną decyzyjną jest *diabetes* i przyjmuje ona 2 klasy: *positive* oraz *negative*

ABALONE - zbiór danych opisujący uchowce (gatunek zwierząt morskich). Pobrany został ze strony: http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data
Zawiera 4177 obiektów oraz 8 atrybutów. Zmienną decyzyjną jest *Type* i przyjmuje ona 3 klasy: *Male*, *Female* oraz *Infant*

W każdym przypadku zbiór danych jest dzielony na trenujący (2/3) oraz testujący (1/3) za pomocą funkcji *sample()*, która pochodzi z pakietu *base*. Na danym zbiorze trenującym oraz przy użyciu wybranego algorytmu budowany jest model, który jest później sprawdzany za pomocą zbioru testowego. Lista parametrów dla danego modelu są tworzona jest pomocy funkcji *confusionMatrix()*, która pochodzi z pakietu *caret*. Z niej można odczytać między innymi takie parametry jak: *Accuracy* (dokładność modelu), a także (dla każdej z klas zmiennej decyzyjnej) *Sensitivity* oraz *Specificity.*

2. INSTALACJA

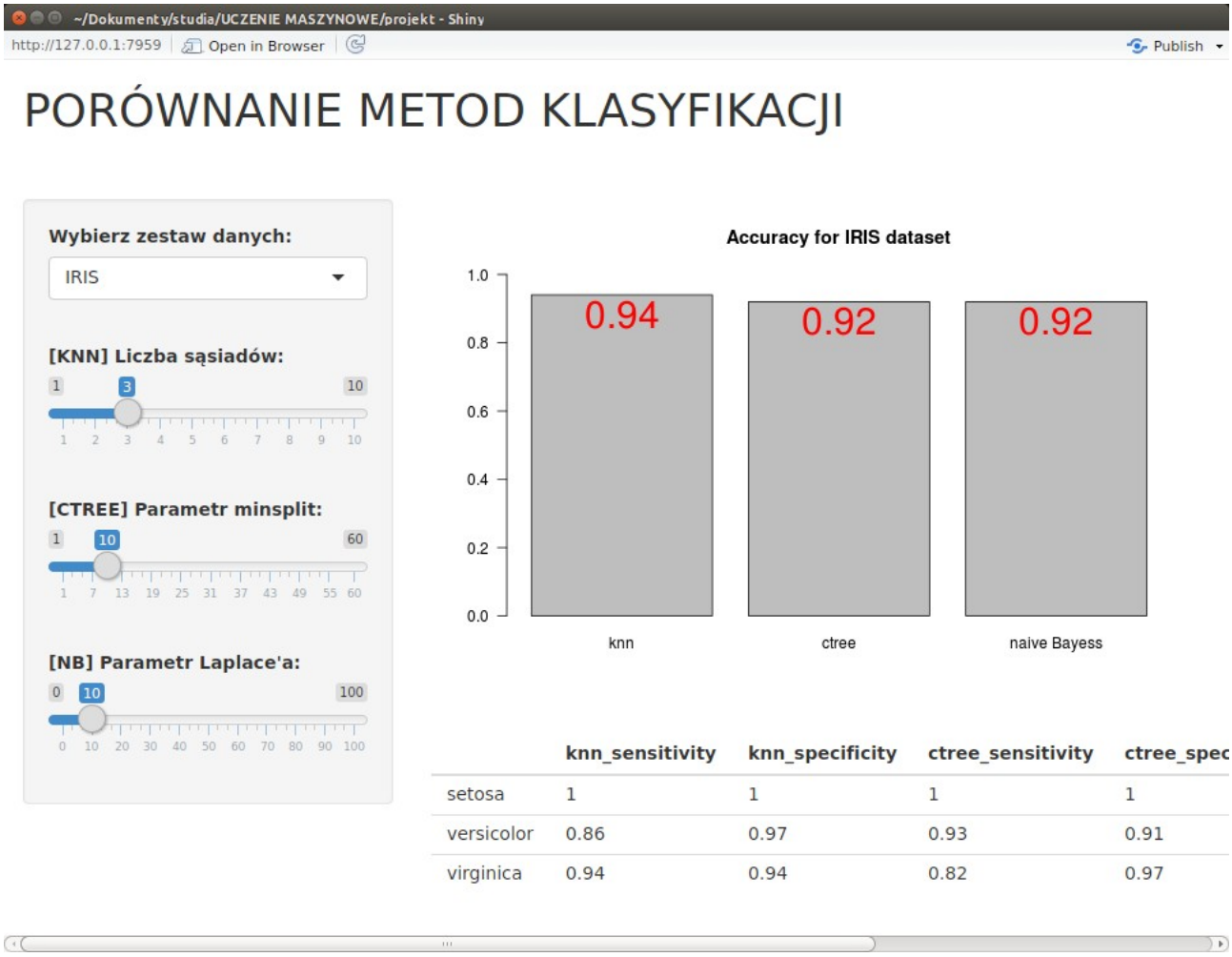
Jeżeli mamy zainstalowane RStudio na naszym komputerze, to wystarczy uruchomić aplikację w tym programie, a następnie kliknąć *Run App*.
Jeżeli mamy tylko zainstalowane środowisko R (bez RStudio), to w terminalu należy ustawić scieżkę dostępu na folder zawierający plik z aplikację, później uruchomić środowisko R, a następnie wpisać komendę:

shiny::runApp()

UWAGA: przed uruchomieniem aplikacji należy mieć zainstalowaną biblioteke *shiny*

W przypadku problemów z automatycznym doinstalowywaniem brakujących pakietów przy użyciu *pacman'a* należy usunąć / wykomentować w pliku aplikacji linie 3, 5, 6, 7, 8 oraz zainstalować brakujące pakiety ręcznie.

3. OPIS INTERFEJSU



Po lewej stronie okna mamy możliwość wyboru zbioru danych. Na wybranym przez użytkownika zbiorze wykonywane są 3 algorytmy klasyfikacji przedstawione w punkcie pierwszym. Za pomocą suwaka można ustawić parametry, również opisane w punkcie pierwszym.

Po prawej stronie okna przedstawiona jest wizualizacja wyników. U góry znajduje się wykres słupkowy, który reprezentuje dokładność (*Accuracy*) każdej metody dla wybranego zbioru danych. Poniżej przedstawiona jest tabela, która pokazuje wartości parametrów *Specificity* oraz *Sensitivity* dla każdej z klas zmiennej decyzyjnej wybranego zbioru danych.

4. INTERPRETACJA WYNIKÓW

IRIS

Dla zbioru IRIS najwyższą dokładność (*Accuracy*) daje model powstały przy użyciu algorytmu *k najbliższych sąsiadów*. Wynosi ona 0.94 (dla k = 3 oraz dla k = 8)

Co ciekawe, w przypadku zbioru IRIS, wartość parametru *minsplit* nie wpływa na dokładność modelu powstałego przy użyciu algorytmu *ctree*. Jest ona stała i wynosi 0.92

Dokładność modelu powstałego przy użyciu algorytmu *naiwnego Bayesa* również wynosi 0.92.

UWAGA: warto odnotować fakt, że dla każdego zbioru danych (IRIS, INDIAN oraz ABALONE) wartość parametru *Laplace'a* nie ma wpływu na dokładność modelu powstałego przy użyciu algorytmu naiwnego Bayesa. Nie ma ona również wpływu na parametry *Sensitivity* oraz *Specificity* dla klas poszczególnych zmiennych decyzyjnych.

INDIAN

Dla zbioru INDIAN dokładność modelu powstałego przy użyciu algorytmu *knn* waha się (w zależności od wartości parametru *k*) od wartości 0.7 do 0.78

Wybrane wartości *Accuracy* (w zależności od parametru *minsplit*) dla modelu powstałego przy użyciu algorytmu *ctree* przedstawia poniższa tabela:

minsplit	1	50	100	150	200	255
Accuracy	0.77	0.77	0.81	0.76	0.78	0.78

Dokładność modelu powstałego przy użyciu algorytmu *naiwnego Bayesa* wynosi 0.78.

UWAGA: tabela (pod wykresem słupkowym w aplikacji) dla zbioru INDIAN zawiera wartości parametrów *Sensitivity* oraz *Specificity* tylko dla jednej z klas zmiennej decyzyjnej. Wynika to z konstrukcji obiektu *confusionMatrix*.

ABALONE

Dla zbioru ABALONE dokładność modelu powstałego przy użyciu algorytmu *knn* waha się (w zależności od wartości parametru *k*) od wartości 0.53 do 0.54

Wybrane wartości *Accuracy* (w zależności od parametru *minsplit*) dla modelu powstałego przy użyciu algorytmu *ctree* przedstawia poniższa tabela:

minsplit	1	250	500	750	1000	1392
Accuracy	0.53	0.53	0.51	0.51	0.51	0.52

Dokładność modelu powstałego przy użyciu algorytmu *naiwnego Bayesa* wynosi 0.51.

5. ZAŁĄCZNIKI

Do niniejszej dokumentacji załączony jest plik *app.R* (zawiera on kod źródłowy aplikacji)