IRD kolokwium 2019/2020, semestr zimowy, wariant C 14.01.2020

Formalności

Rozwiązania wpisuj w pliku o nazwie utworzonej wedle wzorca: "Nazwisko_NrIndeksu.R". Nie musisz przeklejać poleceń, ale zachowaj kolejność zadań i wypisuj wyraźnie ich numery (jako komentarze). Zachowanie struktury pliku przyspieszy sprawdzanie i ogłoszenie wyników.

Przed końcem kolokwium wyślij plik **jako załącznik** do wiadomości mailowej na adres **nosarzewski.aleks@gmail.com**. Tytuł maila powinien zawierać Twój numer indeksu.

Rozwiązania wklejone w treści maila (zamiast załączenia jako plik .R) albo wysłane po czasie, nie będą sprawdzane. Kolokwium zostanie sprawdzone przez uruchamianie kodu linijka po linijce. Upewnij się, że fragmenty kodu są we właściwej kolejności oraz cały skrypt wykonuje się poprawnie.

Zadanie 1 (10 pkt)

Napisz funkcję `persistence`, która będzie liczyła tzw. persistence (trwałość) liczby (`link`). Jeżeli dla danej liczby przemnożymy wszystkie jej cyfry przez siebie, następnie dla wyniku otrzymanego powtórzymy operację i będziemy tę czynność powtarzać tak długo, dopóki nie dostaniemy liczby jednocyfrowej (mniejszej od 10), to liczbę takich operacji nazywamy trwałością liczby. Przykładowo dla liczby 9999:

```
9 * 9 * 9 * 9 = 6561
6 * 5 * 6 * 1 = 180
1 * 8 * 0 = 0
```

Zatem trwałość liczby 9999 wynosi 3.

PODPOWIEDŹ: Aby rozbić liczbę na poszczególne cyfry mozesz zastosować poniższe polecenie. Otrzymasz wtedy wektor zawierający poszczególne cyfry liczby:

```
liczba <- 9876
as.numeric(unlist(strsplit(as.character(liczba), split = NULL)))
```

```
## [1] 9 8 7 6
```

Twoja funkcja powinna działać w następujący sposób:

```
persistence(9) = 1
persistence(10) = 1
persistence(277) = 4
persistence(769) = 5
persistence(277777788888899) = 11
```

Zaaplikuj funkcję na swoim numerze albumu.

Zadanie 2 (10 pkt)

Podpowiedź do zadania - do danych z pakietu `lattice` można dostać się w następujący sposób:

```
library(lattice)
barley <- barley
```

Na podstawie danych `barley` z pakietu `lattice`:

- Z wykorzystaniem operatora pipeline (`%>%`) z pakietu `dplyr`, oblicz następujące statystyki dla zbiorów jęczmienia (*yield*): minimum, medianę, średnią, odchylenie standardowe oraz maksimum w zależności od gatunku jęczmienia (*variety*). Wyniki posortuj malejąco według wartości średniego zbioru. Która odmiana jęczmienia dawała średnio najwyższe zbiory (odpowiedź napisz jako komentarz)?
- Narysuj wykres (wykorzystując pakiet `ggplot2`) taki sam jak w Załączniku 1.

Zadanie 3 (20 pkt)

Podpowiedź do zadania - do danych z pakietu `mlbench` można dostać się w następujący sposób:

```
library(mlbench)
data("PimaIndiansDiabetes")
```

Na podstawie danych `PimaIndiansDiabetes` z pakietu `mlbench`:

a) Wczytaj dane oraz podziel zbiór na uczący i testowy w proporcji 75% do 25%. (Ziarno losowe ustaw na swój numer indeksu).
b) Zbuduj dwa modele prognozujące zmienną `diabetes`:
   - drzewo klasyfikacyjne z wszystkimi zmiennymi objaśniającymi przy domyślnych ustawieniach,
   - las losowy z wszystkimi zmiennymi objaśniającymi (ustaw liczbę wykorzystanych drzew na 200).
c) Która ze zmiennych objaśniających ma największy wpływ na zmienną prognozowaną (odpowiedź napisz jako komentarz)?
d) Na podstawie zbioru testowego, policz `Accuracy`, `Recall` oraz `False Negative Rate (FNR)`. Oceń, który model lepiej prognozuje zmienną `diabetes` ze względu na `FNR` (odpowiedź napisz jako komentarz)?
e) Dla lepszego modelu policz AUC (napisz jego wartość jako komentarz) oraz narysuj ROC.

Zadanie 4 (10p)

Podpowiedź do zadania - do danych z pakietu `AppliedPredictiveModeling` można dostać się w następujący sposób:

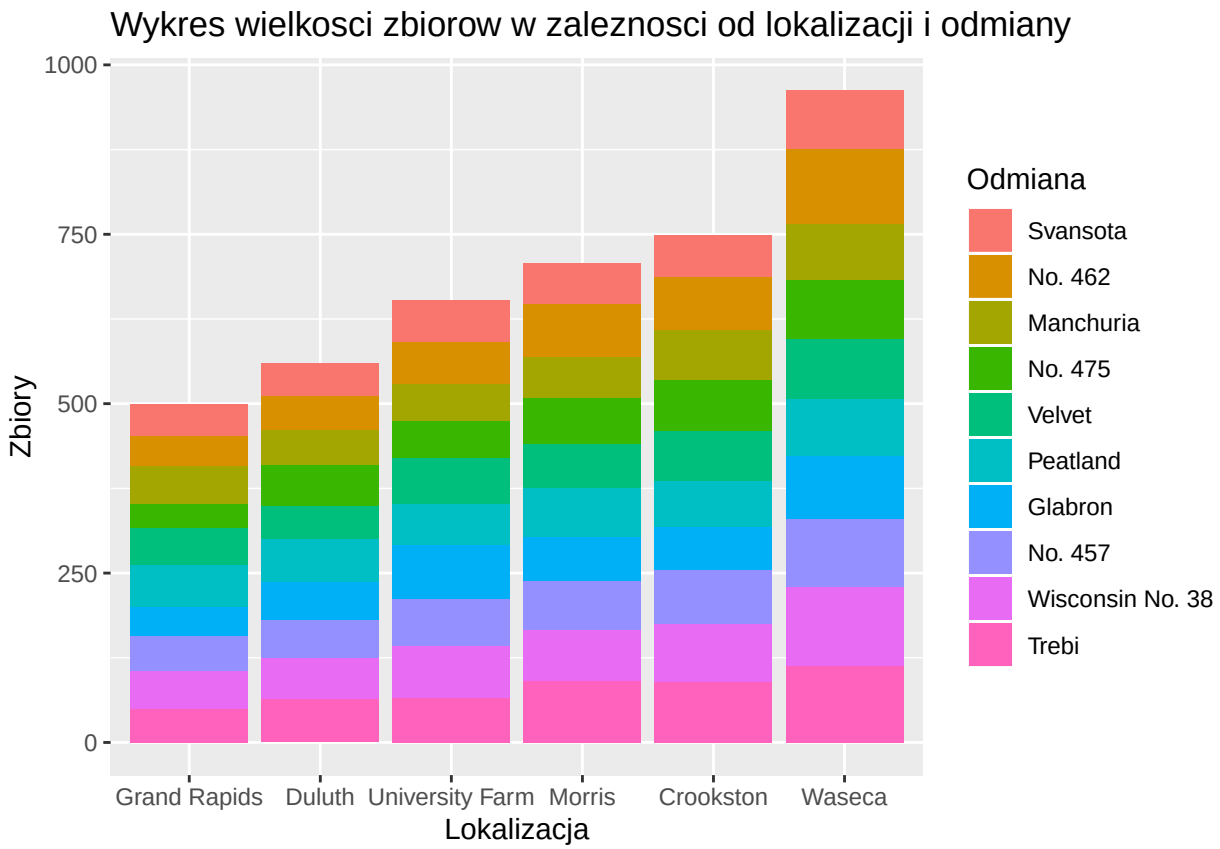
```
#install.packages("AppliedPredictiveModeling")
library(AppliedPredictiveModeling)
data(abalone)
```

Na podstawie danych `abalone` z pakietu `AppliedPredictiveModeling`:

a) Podziel zbiór na uczący i testowy w proporcji 80% do 20% (ziarno losowe ustaw na swój numer indeksu).
b) Na zbiorze uczącym zbuduj modele prognozującą zmienną `Rings`:
   - regresji liniowej z wszystkimi zmiennymi jako zmiennymi objaśniającymi,
   - drzewa regresyjnego z wszystkimi zmiennymi jako zmiennymi objaśniającymi (ustaw minimalną liczbę obserwacji w liściu końcowym (terminal node) na 50).
c) Podaj słownie jedną przykładową regułę otrzymaną z drzewa (jako komentarz, pamiętaj o różnicy w znaczeniu prognozy między drzewem klasyfikacyjnym a regresyjnym).
d) Na podstawie zbioru testowego oblicz:
   - Mean Squared Error (MSE),
   - Mean Absolute Error (MAE).

   Który model jest lepszy (odpowiedź napisz jako komentarz i uzasadnij)?

Załącznik 1


Wykres wielkosci zbiorow w zaleznosci od lokalizacji i odmiany
Zbiory
1000
750
500
250
0
Grand Rapids
Duluth
University Farm
Morris
Crookston
Waseca
Lokalizacja
Odmiana
Svansota
No. 462
Manchuria
No. 475
Velvet
Peatland
Glabron
No. 457
Wisconsin No. 38
Trebi