Tytuł:
DOPASOWANIE MODELU DO DANYCH ZA POMOCĄ REGRESJI LOGISTYCZNEJ

Zespół:
Oktawia Hankus, Zuzanna Nogala

Motywacja:
Stworzenie wygodnego narzędzia, które sprawnie oceni, który model będący regresją logistyczną będzie najlepiej dopasowany do danych przekazanych przez użytkownika. Użytkownik na podstawie wybranych przez siebie predyktorów i zmiennej odpowiedzi otrzyma w aplikacji R za pomocą pakietu „projektROR" wstępną analizę danych oraz stworzyć odpowiedni model na jej podstawie. Daje możliwość wykonania skomplikowanych obliczeń za jednym wywołaniem funkcji: AIC, BIC, walidacja krzyżowa. Pozwala także zwizualizować predykcję, kiedy zmienna Y zależy od jednego regresora. Pakiet zaopatrzony jest w dwa przykładowe zbiory danych, które pozwolą poznać jego możliwości, bez konieczności wgrywania własnych danych.

Dane:
Pakiet zawiera dwa zbiory danych. Pierwszy „citrus" zawiera 10000 obserwacji o tym czy dany owoc jest grejpfrutem czy pomarańczą za pomocą wartość RGB koloru, wagi czy średnicy owocu. Drugi zbiór „creditData" zawiera Informacje o statusie spłaty kredytu 20762 unikatowych klientach banku oraz pozostałych 19 cech, między innymi:

- Płeć,
- Liczba dzieci
- Czy posiada własny samochód, mieszkania
- Wykształcenie
- Zawód
- Wiek

itd.

- **Format danych:** plik .csv
- **Źródła:** citrus - https://www.kaggle.com/datasets/joshmcadams/oranges-vs-grapefruit
  creditData - https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

Narzędzia:

Kluczowe pakiety:
data.table
ggplot2,
kableExtra,

Inne:
Github

Planowanie funkcjonalności:

Konieczne funkcjonalności pakietu:
1. Wstępna wizualizacja danych. (Wykresy pudełkowe przedstawiające zależności pomiędzy predyktorami a zmienną objaśniającą)
2. Stworzenie modelu.
3. Porównanie modeli za pomocą testu zgodności oparty na statystyce Deviance,
4. Analiza modelu na podstawie walidacji krzyżowej.
5. Porównywanie modeli na bazie krzywych ROC i ich wartości statystyki AUC.

6. Macierz błędów (ang. Error/Confusion Matrix)
7. Obliczanie AIC i BIC dla wszystkich możliwych modeli przy małej liczbie predyktorów.
8. Predykcja nowych wartości.

Opcjonalne funkcjonalności:
1. Krzywa predykcji prawdopodobieństwa dla modelu z jedną zmienną.
2. Graficzne przedstawienie macierz kowariancji w postaci wykresu „heatmap”.