STATYSTYKA II - PROJEKT

Michał Maj

3/29/2021

Celem projektu jest analiza danych dotyczących udaru mózgu, znajdujących się w pliku `healthcare-dataset-stroke-data.cs`
Wyniki należy przedstawić w formie 20-minutowej prezentacji (ostatnie 2-3 wykłady) oraz raportu (RMarkdown, Jupyter, etc.) zawierającego kod programu jak i opisy dokonywanych decyzji. Analiza powinna zawierać co najmniej następujące punkty:

1. Czyszczenie danych (usuwanie/inputacja braków danych, naprawa błędów, transformacje danych, rozwiązanie problemu wartości odstających)
2. Eksploracyjna analiza danych
3. Zamodelowanie zmiennej `stroke` na podstawie pozostałych zmiennych. Minimum 3 modele z czego jednym z nich musi być Random Forest.
4. Ewaluacja na zbiorze testowym (wybór modelu i metryk z uzasadnieniem)

Możliwe jest rozwiązaniem w języku R, Python lub SAS