Ćwiczenie laboratoryjne

Praca z danymi medycznymi: import, przetwarzanie i analiza zbiorów danych pacjentów

Katedra Informatyki i Automatyki

1 Cel ćwiczenia

Celem ćwiczenia jest zapoznanie studentów z procesem pracy z danymi medycznymi, obejmującym:

- import zbiorów danych pacjentów z plików w popularnych formatach,
- wstępne przetwarzanie danych (czyszczenie, uzupełnianie braków, standaryzacja),
- analizę statystyczną i wizualizację informacji,
- przygotowanie danych do dalszego wykorzystania w systemach wspomagania decyzji medycznych.

2 Wprowadzenie teoretyczne

Dane medyczne stanowią szczególną kategorię danych wrażliwych. Mogą pochodzić z systemów elektronicznej dokumentacji medycznej (EHR), badań klinicznych, urządzeń monitorujących lub ankiet pacjentów. Ich analiza umożliwia m.in.:

- monitorowanie stanu zdrowia pacjentów,
- wczesne wykrywanie chorób,
- ocenę skuteczności terapii,
- wspomaganie decyzji klinicznych.

Typowe problemy związane z danymi medycznymi:

- brakujące wartości i błędy pomiarowe,
- niejednorodne formaty danych (CSV, Excel, XML, DICOM),
- konieczność anonimizacji danych pacjentów,
- wysoka zmienność i wielowymiarowość danych.

3 Stosowane technologie i biblioteki

W ćwiczeniu wykorzystane zostaną następujące narzędzia i biblioteki:

Python – język programowania szeroko stosowany w analizie danych.

Jupyter Notebook – środowisko interaktywne do eksperymentów i prezentacji wyników.

Pandas – biblioteka do pracy z tabelarycznymi danymi (import, transformacje, grupowanie).

NumPy – biblioteka do obliczeń numerycznych i macierzowych.

Matplotlib, Seaborn – narzędzia do wizualizacji danych.

Scikit-learn – biblioteka uczenia maszynowego wspierająca analizę i modelowanie danych.

4 Przebieg ćwiczenia

Etap 1: Import danych

- Zaimportuj dane pacjentów z pliku CSV lub Excel.
- Zapoznaj się ze strukturą zbioru (nagłówki, liczba rekordów, typy danych).

Etap 2: Wstępne przetwarzanie danych

- Sprawdź obecność brakujących danych i zastosuj wybraną metodę ich uzupełnienia (np. średnia, mediana, usunięcie rekordu).
- Dokonaj standaryzacji lub normalizacji wybranych zmiennych (np. ciśnienie, BMI).
- Zmień typy danych w razie potrzeby (np. daty, kategorie).

Etap 3: Analiza i wizualizacja

- Oblicz podstawowe statystyki opisowe (średnia, mediana, odchylenie standardowe).
- Wykonaj wizualizacje: histogramy, wykresy pudełkowe, korelacje między zmiennymi.
- Zidentyfikuj potencjalne zależności w danych (np. wpływ wieku na ciśnienie).

Etap 4: Przygotowanie do modelowania

- Podziel dane na zbiór treningowy i testowy.
- Zastosuj prostą metodę klasyfikacyjną (np. regresję logistyczną) dla wybranego problemu.

5 Warianty zadań dla studentów

1. Importuj dane pacjentów i przeprowadź analizę statystyczną zmiennych.
2. Uzupełnij brakujące dane w zbiorze badań i sprawdź zależność.
3. Przygotuj wizualizację pokazującą częstość występowania różnych kategorii w populacji pacjentów.
4. Zbuduj prosty model predykcyjny.
5. Porównaj skuteczność dwóch metod normalizacji danych w kontekście poprawy jakości klasyfikacji.

Linki do zbiorów danych

1. klasyfikacja choroby Parkinsona `https://www.kaggle.com/dipayanbiswas/parkinsons-diseas`
   `select=pd_speech_features.csv`
2. smoking patients `https://www.kaggle.com/thomaskonstantin/cpg-values-of-smoking-and-`
3. Powikłania zawału mięśnia sercowego `https://www.kaggle.com/rafatashrafjoy/`
   `myocardial-infarction-complications`
4. Sygnały kardiotokografii `https://www.kaggle.com/sohelranaccselab/biomedical-cardiotoc`
5. Badania pH `https://www.kaggle.com/zfturbo/measurements-of-urine-ph`
6. Analiza cukrzycy `https://www.kaggle.com/veerukhannan/diabetes`
7. Choroba Alzheimera `https://www.kaggle.com/madhucharan/alzheimersdisease5classdatase`
8. Prostate cancer `https://www.kaggle.com/ashrafalsinglawi/prostate-cancer-survival-da`
9. klasyfikacja choroby Parkinsona `https://www.kaggle.com/dipayanbiswas/parkinsons-diseas`
   `select=pd_speech_features.csv`
10. smoking patients `https://www.kaggle.com/thomaskonstantin/cpg-values-of-smoking-and-`
11. Powikłania zawału mięśnia sercowego `https://www.kaggle.com/rafatashrafjoy/`
    `myocardial-infarction-complications`

12. Sygnały kardiotokografii https://www.kaggle.com/sohelranaccselab/biomedical-cardiotoc

13. Badania pH https://www.kaggle.com/zfturbo/measurements-of-urine-ph

14. Analiza cukrzycy https://www.kaggle.com/veerukhannan/diabetes

15. Choroba Alzheimera https://www.kaggle.com/madhucharan/alzheimersdisease5classdatase

16. Prostate cancer https://www.kaggle.com/ashrafalsinglawi/prostate-cancer-survival-da

6 Literatura i źródła

- Wes McKinney, *Python for Data Analysis*, O'Reilly Media, 2017.
- Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, 2022.
- Dokumentacja biblioteki Pandas: https://pandas.pydata.org/
- Dokumentacja biblioteki Scikit-learn: https://scikit-learn.org/