**Zaawansowane metody i techniki analizy danych**

**Raport z realizacji projektu**

Celem projektu było stworzenie klasyfikatorów umożliwiających ocenę funkcjonowania pracy serca na podstawie ilościowej analizy sygnału EKG. Projekt dostępny w repozytorium https://github.com/MartaPiatek/ZMITAD

Zbiór danych, zwierający zapisy sygnałów elektrokardiograficznych uzyskane z 14 odprowadzeń, pobrano ze bazy PhysioNet [1]. Zawiera on 549 zapisów z badań 290 pacjentów (52 osoby zdrowe) w wieku 17-87 lat (209 mężczyzn i 81 kobiet). Sygnały były próbkowane z częstotliwością 1000 Hz.

Ze względu na różną długość sygnałów do dalszej analizy wybrano jedynie 245 zapisów,   
w tym 56 pochodzących od osób zdrowych oraz 189 od osób z zaburzeniami pracy serca.

W celu uproszczenia analizy wykorzystywano tylko sygnał zarejestrowany na I odprowadzeniu EKG.

Pierwszym etapem realizacji projektu była implementacja algorytmu segmentacji sygnału EKG. Wykorzystano algorytm opracowany przez Pan & Tompkins [2].

Następnie zaimplementowano funkcje obliczające parametry charakteryzujące pracę serca:

* amplitudy i czasy trwania załamków P,Q,R,S,T,
* czas trwania zespołu QRS,
* czas trwania odcinków PQ i ST
* czas trwania odstępów PQ, QT
* częstość pracy serca i czas trwania odcinka RR

Dzięki temu dla każdego pacjenta otrzymano 17 cech charakteryzujących pracę jego serca.

Wartości wyznaczonych parametrów dla całej grupy pacjentów zapisano w pliku tekstowym.

W celu określenia statusu pacjenta (zdrowy/z zaburzeniami pracy serca) zastosowano następujące metody klasyfikacji:

* k-means
* drzewo decyzyjne
* naiwna metoda Bayesa
* SVM
* k-najbliższych sąsiadów

Ze względu na duże różnice w licznościach osób zdrowych i chorych, zbiór danych poddany klasyfikacji został okrojony tak, by obie grupy były równoliczne. Zatem spośród 189 elektrokardiogramów pacjentów z zaburzeniami pracy serca wybrano jedynie 56 zapisów   
i dołączono do puli sygnałów osób zdrowych.

Poniżej zamieszczono wyniki klasyfikacji.

* **k-means**

Zaimplementowano algorytm k-means grupujący pacjentów na 2 klasy (zdrowy/z zaburzeniami pracy serca) dla wszystkich 17 cech.

Dokładność klasyfikacji tego algorytmu dla:

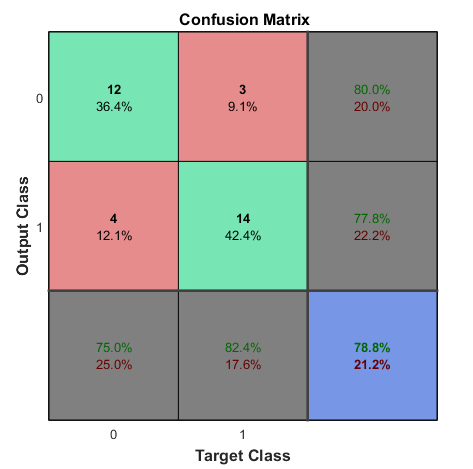
* + pacjentów zdrowych wyniosła: 20%
  + pacjentów z zaburzeniami pracy serca wyniosła: 91%
  + wszystkich pacjentów: 56%
* **drzewo decyzyjne**

Do stworzenia drzewa wykorzystano funkcję MATLAB-a **ClassificationTree.**

Wyznaczono macierz błędów klasyfikacji (rysunek 1). Klasa ‘0’ oznacza osoby z zaburzeniami pracy serca, natomiast ‘1’ to osoby zdrowe. Do zbioru treningowego przydzielono 70% danych, pozostałe 30% stanowi zbiór testowy.

Dokładność klasyfikacji tego algorytmu dla:

* + pacjentów zdrowych wyniosła: 80%
  + pacjentów z zaburzeniami pracy serca wyniosła: 77,8%
  + wszystkich pacjentów: 78,8%



Rysunek 1. Macierz błędów dla klasyfikacji z wykorzystaniem drzewa decyzyjnego

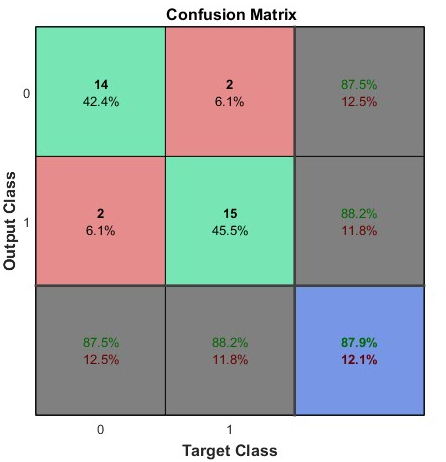
* **naiwny klasyfikator Bayesa**

Dla klasyfikacji danych metodą Bayesa wykorzystano funkcję **fitcnb**.

Na rysunku 2 zaprezentowano macierz pomyłek klasyfikacji. Sposób oznaczenia klas oraz podział na zbiór treningowy i testowy jest taki sam jak w przypadku drzewa decyzyjnego.

Dokładność klasyfikacji tego algorytmu dla:

* + pacjentów zdrowych wyniosła: 88,2%
  + pacjentów z zaburzeniami pracy serca wyniosła: 87,5%
  + wszystkich pacjentów: 87,9%



Rysunek 2. Macierz błędów dla klasyfikacji z wykorzystaniem naiwnego klasyfikatora Bayesa

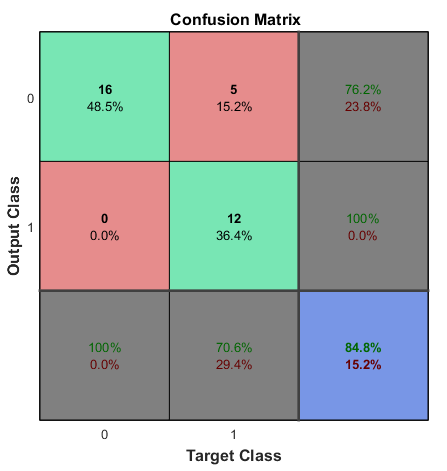
* **SVM**

Dla klasyfikacji danych metodą SVM wykorzystano funkcję **fitcsvm**.

Na rysunku 3 zaprezentowano macierz pomyłek klasyfikacji. Sposób oznaczenia klas oraz podział na zbiór treningowy i testowy jest taki sam jak w przypadku drzewa decyzyjnego.

Dokładność klasyfikacji tego algorytmu dla:

* + pacjentów zdrowych wyniosła: 100%
  + pacjentów z zaburzeniami pracy serca wyniosła: 76,2%
  + wszystkich pacjentów: 84,8%



Rysunek 3. Macierz błędów dla klasyfikacji z wykorzystaniem SVM

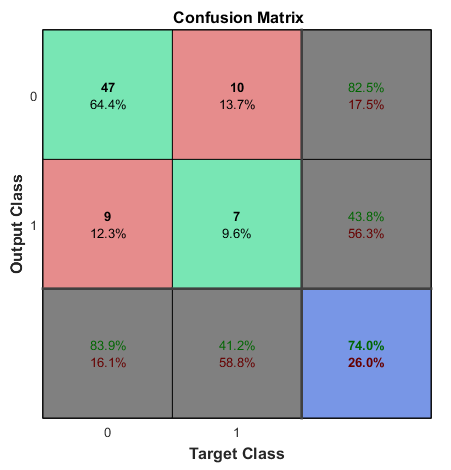
* **najbliższych sąsiadów**

Dla klasyfikacji danych metodą najbliższych sąsiadów wykorzystano funkcję **fitcknn**.

Na rysunku 4 zaprezentowano macierz pomyłek klasyfikacji. Sposób oznaczenia klas oraz podział na zbiór treningowy i testowy jest taki sam jak w przypadku drzewa decyzyjnego. Jako liczbę najbliższych sąsiadów ustawiono wartość 8.

Dokładność klasyfikacji tego algorytmu dla:

* + pacjentów zdrowych wyniosła: 43,8%
  + pacjentów z zaburzeniami pracy serca wyniosła: 82,5%
  + wszystkich pacjentów74%



Rysunek 4. Macierz błędów dla klasyfikacji z wykorzystaniem metody najbliższych sąsiadów

**Wnioski:**

Metodą k-means bardzo dobrze sklasyfikowano osoby chore. Ponad 90% skuteczność pozwala wywnioskować, iż mimo swej prostoty algorytm dobrze wychwycił pacjentów, u których wystąpiły odchylenia od normy. Z kolei zaledwie 20% skuteczność dla osób zdrowych, może wskazywać, iż zastosowane parametry nie są wystarczające do jednoznacznej diagnozy. Należy podkreślić fakt, iż parametry wyznaczano jedynie z jednego odprowadzenia EKG. Być może, rozszerzenie analizy o większą liczbę kanałów skutkowałoby zwiększeniem skuteczności działania klasyfikatora. Jednakże ze względu na dziedzinę zastosowania aplikacji, zdecydowanie lepszą sytuacją jest zakwalifikowanie pacjenta jako fałszywie chorego niż fałszywie zdrowego.

Zdecydowanie lepszą klasyfikację pacjentów uzyskano metodą drzewa decyzyjnego. Skuteczność algorytmu wyniosła blisko 80% dla obu grup osób.

Wyniki klasyfikacji uzyskane dla metody SVM i naiwnego klasyfikatora Bayesa są podobne. Średnia skuteczność całkowitej klasyfikacji powyżej 80% wskazuje na bardzo dobrą segregację pacjentów.

Ostatnia przebadana metoda – najbliższych sąsiadów, pozwala uzyskać niższe wskaźniki skuteczności. Dla pacjentów chorych klasyfikacja nadal jest bardzo dokładna – skuteczność ponad 80%. Jednak w przypadku pacjentów zdrowych skuteczność wynosi ok. 44%.

Podsumowując, najskuteczniejszymi metodami klasyfikacji okazały się: drzewo decyzyjne, SVM i naiwny klasyfikator Bayesa. Wysokie współczynniki skuteczności mogą sugerować, iż taki klasyfikator z powodzeniem mógłby być wykorzystywany w praktyce klinicznej jako narzędzie wspomagające pracę lekarza diagnozującego choroby serca. Oczywiście bardziej wiarygodnym wynikiem klasyfikacji byłaby analiza sygnałów pochodzących z większej liczby odprowadzeń EKG.

**Bibliografia**

1. Baza danych PhysioNet <https://www.physionet.org/physiobank/database/ptbdb/>
2. J. Pan, W. Tompkins „A real-time QRS detection algorithm”, Transactions on Biomedical Engineering, vol 32, 1985