Przewidywanie przeżycia pacjentów z niewydolnością serca na podstawie serum kreatyniny i frakcji wyrzutowej

1) **Artykuł i źródło danych**

- **Projekt został zrealizowany w oparciu o artykuł:** Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* **20**, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5 znaleziony na stronie: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone | BMC Medical Informatics and Decision Making | Full Text (biomedcentral.com) (open access)
- **Źródło i opis danych:** Dane do projektu zostały pobrane ze strony UCI – Machine Learning repository. W źródłach literaturowych artykułu znajduje się bezpośredni link do źródła : UCI Machine Learning Repository: Heart failure clinical records Data Set (open access)

2) **Opis zestawu danych**

Zestaw danych zawiera dokumentację medyczną **299 pacjentów** z niewydolnością serca, zebraną w okresie obserwacji. Każdy profil pacjenta ma **13 cech klinicznych**. Dane zostały zebrane w 2015 roku

1. **(age) wiek**: wiek pacjenta (w latach)
2. **(anamea) niedokrwistość**: zmniejszenie liczby czerwonych krwinek lub hemoglobiny (logiczna)
3. **(high blood pressure) wysokie ciśnienie krwi**: jeśli pacjent ma nadciśnienie (logiczne)
4. **(creatinine) fosfokinaza kreatyniny (CPK)**: poziom enzymu CPK we krwi (mcg/l)
5. **(diabetes) cukrzyca:** jeśli pacjent ma cukrzycę (boolean)
6. **(ejection fraction) frakcja wyrzutowa**: procent krwi opuszczającej cerce przy każdym skurczu (w procentach)
7. **(platelets) płytki krwi**: płytki krwi we krwi pacjenta (kilopłytki/ml)
8. **(sex) płeć**: kobieta lub mężczyzna (binarne)
9. **(serum creatinine) kreatynina w surowicy**: poziom kreatyniny w surowicy krwi (mg/dl)
10. **(serum sodium) sód w surowicy**: poziom sodu w surowicy krwi (mEq/L)
11. **(smoking) palenie**: czy pacjent pali, czy nie (boolean)
12. **(time) czas**: okres obserwacji (dni)
13. **([target] death) [docelowe] zdarzenie śmierci**: jeśli pacjent zmarł w okresie obserwacji (boolean)

3) **Cel projektu:**
(analiza datasetu) Celem projektu jest: Podwierdzenie dwóch cech będących w największej korelacji ze zdarzeniem śmierci u pacjentów (od strony analizy danych)
(Machine Learning part) Celem projektu zastosowanie znalezienie z pośród zaproponowanych przez autorów klasyfikatorów, najlepszej metody do oceny przeżycia.

4) Zastosowane metody (zastosowano algorytmy uczenia maszynowego zaproponowane w artykule badawczym):
    - Regresja logistyczna

- SVM
- Drzewo Decyzyjne
- Gradient Boosting
- KNN – K-najbliżsi sąsiedzi
- Sztuczna sieć neuronowa

5) Wyniki i porównanie z danymi otrzymanymi w artykule.

Korelacja zmiennych ze zdarzeniem śmierci – w artykule autorzy wyodrębnili dwie najbardziej skorelowane zmienne ze zmienną „Death Event" – zdarzenie śmierci

Jeśli chodzi o analizę przeżycia, to najlepsza metoda uczenia maszynowego zastosowana do analizy (z najwyższym Accuracy, to Drzewa, to las losowy, nieco mniejsze Accuracy miał las losowy i drzewa decyzyjne. Najsłabsza metoda do analizy czynników względem czynnika śmierci, to SVM i K-neighbours.

6) WNIOSKI:

Przeprowadzone obliczenia potwierdzają podane w publikacji informacje. W publikacji w zastosowaniu niektórych metod dzielono zbiór testowy i walidacyjny odrobinę inaczej (np. 80/20). Ja ze względu na to, że chciałam przetestować podane metody na wybranych zbiorach, zdecydowałam się na jeden podział (z zastosowaniem SMOTE – ze względu na niezbalansowany rozkład danych. Obok zastosowanych algorytmów uczenia maszynowego, autorzy jednocześnie implementowali metody w języku R – tę część pracy pominęłam. W związku ze zmianami (np. jeden podział datasetu) wyniki nie są identyczne, ale zbliżone. Np. Las losowy u autorów, to najlepszy możliwy algorytm przewidywania przeżycia – w przypadku moich obliczeń bardzo wysokie accuracy wyszło również w SVM i drzewach decyzyjnych.