Raport do projektu zaliczeniowego WdAD

Celem analizy danych przeprowadzonej w naszym projekcie jest stworzenie modelu przewidującego śmiertelność pacjenta spowodowaną niewydolnością serca. Choroby sercowo-naczyniowe są główną przyczyną zgonów na całym świecie, pochłaniając około 17,9 miliona istnień ludzkich każdego roku, co stanowi 31% wszystkich zgonów na świecie. Niewydolność serca jest częstym zdarzeniem spowodowanym przez choroby sercowo-naczyniowe, nasz zbiór danych zawiera 12 cech, które można wykorzystać do przewidywania śmiertelności z powodu niewydolności serca.Większości chorób sercowo-naczyniowych można zapobiegać poprzez przeciwdziałanie behawioralnym czynnikom ryzyka, takim jak palenie tytoniu, niezdrowa dieta i otyłość, brak aktywności fizycznej i szkodliwe spożywanie alkoholu przy użyciu strategii obejmujących całą populację. Osoby z chorobami sercowo-naczyniowymi lub osoby z wysokim ryzykiem sercowo-naczyniowym wymagają wczesnego diagnozowania, w którym pomocny może być model uczenia maszynowego.

Krok 1. Zbieranie danych

Zbiór danych użyty w projekcie pochodzi ze strony kaggle i można go odnaleźć pod linkiem https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data?datasetId=727551&language=R

Zbiór zawiera 12 cech:

§ age: wiek badanego

§ anaemia: występowanie anemii (zmniejszenie liczby czerwonych krwinek lub hemoglobiny)

§ creatinine_phosphokinase: poziom enzymu CPK we krwi (mcg/L)

§ diabetes: występowanie u pacjenta cukrzycy

§ ejection_fraction: procent krwi opuszczającej serce przy każdym skurczu

§ high_blood_pressure: czy pacjent ma nadciśnienie

§ platelets: płytki krwi we krwi (kilopłytki/ml)

§ serum_creatinine: poziom kreatyniny we krwi (mg/dl)

§ serum_sodium: poziom sodu we krwi (mEq/L)

§ sex: kobieta lub mężczyzna

§ smoking: czy pacjent pali czy nie

§ czas: okres obserwacji

§ DEATH_EVENT: Czy pacjent zmarł w okresie obserwacji

Krok 2. Eksploracja i przygotowanie danych

Zbiór danych nie zawiera żadnych brakujących wartości, więc nie musimy wykonywać żadnego przypisywania.

Dla lepszego zrozumienia danych przytoczymy kilka statystyk:

Ejection Fraction: Normalna frakcja wyrzutowa wynosi około 50% do 75%, zgodnie z American Heart Association. Graniczna frakcja wyrzutowa może wynosić od 41% do 50%.

Serum Creatinine: Dla dorosłych mężczyzn od 0,74 do 1,35 mg/dl (od 65,4 do 119,3 mikromoli/l) Dla dorosłych kobiet od 0,59 do 1,04 mg/dl (od 52,2 do 91,9 mikromoli/l).

Platelets: Normalna liczba płytek krwi wynosi od 150 000 do 450 000 płytek krwi na mikrolitr krwi.

Serum Sodium: Prawidłowy poziom sodu we krwi wynosi od 135 do 145 miliekwiwalentów na litr (mEq/L).

CreatininePhosphkinase: od 10 do 120 mikrogramów na litr (mcg/L)

Krok 3. Budowa modelu

Metoda k-NN przewiduje klasę danego przykładu na podstawie k najbliższych obserwacji ze zbioru uczącego. Zmienna k jest parametrem tej metody i przyjmuje wartość z zbioru liczb naturalnych dodatnich. Dla ustalonego k, dla każdego przykładu ze zbioru testowego algorytm znajduje k przykładów ze zbioru uczącego, które są najbliższe pod względem podobieństwa do danego przykładu ze zbioru testowego.

Używając funkcji knn(), przeprowadziliśmy klasyfikację przykładów ze zbioru testowego. Ponieważ dane uczące składają się z 229 przykładów, jako pierwsze wypróbowaliśmy k=15. Jednak po zbadaniu wielu możliwości k = 23 daje najlepsze wyniki ze skutecznością 82%.

Krok 4. Ocena modelu

W naszym modelu uzyskaliśmy skuteczność 82%. Łącznie mamy 13 błędnych przewidywań.

Dla małych wartości parametru k (k=5,6,7) wyniki fałszywie negatywne zawierają się w przedziale [2,4], jednak wyników fałszywie pozytywnych jest dosyć dużo . Dla kolejnych

wartości parametru k liczba wyników fałszywie negatywnych utrzymuje się na stałym poziomie, natomiast maleje liczba wyników fałszywie pozytywnych.

W przypadku naszych danych najbardziej zależy nam na zminimalizowaniu wyników fałszywie negatywnych, jednak ich liczba nie zmienia się znacząco w zależności od wybranego k.

Ostatecznie dla k=23 otrzymujemy wynik (3,9), który jest najlepszym z otrzymanych. Predykcje uzyskane z zastosowaniem naszego modelu są zatem zadowalające.

Krok 5. Dopracowanie modelu

Aby zweryfikować jakość zbudowanego modelu stosowaliśmy różne wartości parametru k. Tym sposobem sprawdzaliśmy jego zachowanie przy znormalizowanych danych do przedziału [0,1] metodą min-max. Dodatkowo sprawdziliśmy zachowanie modelu dla k = 1, w wyniku którego, dla tego samego co wcześniej zbioru testowego otrzymaliśmy 4 wyniki fałszywie negatywne. Postawiony problem został rozwiązany w znacznym zakresie. Otrzymaliśmy satysfakcjonujące wyniki na temat śmiertelności pacjenta z powodu niewydolności serca. Zastosowaliśmy funkcję, która losowo permutuje obserwacje i przetestowaliśmy model jeszcze kilkukrotnie.

Problem predykcji śmiertelności wśród pacjentów został w znacznej mierze rozwiązany. Na podstawie kluczowych cech, z wykorzystaniem metody k-NN jesteśmy w stanie z wysoką skutecznością przewidzieć rokowania dla pacjenta. Model będzie się dobrze generalizował.

25.06.2023r. Karolina Grzech, Wioleta Gackowiec, Krystian Bułat