10.12.2018

Błażej Kucman

ind. 238228

gr.1

Inteligencja Obliczeniowa – Projekt 2

Zgłębianie danych

Tematem zadania było przećwiczenie poznanych na laboratoriach technik zgłębiania danych na wybranym zbiorze danych (najlepiej z kolumną „class” z dwoma możliwymi wartościami).

Do zadania została wybrana baza z wyniki i danymi osób z problemami chorób serca. Na ich podstawie chcemy się dowiedzieć przy jakich wynikach oraz danych osoby jest możliwa choroba serca.

Baza składa się z kolumn:

Skrót ST – częstoskurcz zatokowy

age - wiek w latach,

sex – płeć(1=mężczyzna, 0= kobieta),

cp – rodzaj bólu w klatce piersiowej (wartości 1-4)

1. Typowa angina
2. Nietypowa angina
3. Nie anginowy ból
4. Bezobjawowy

trestbps - spoczynkowe ciśnienie krwi (w mm Hg przy przyjęciu do szpitala)

chol - poziom cholesterolu w surowicy w mg/dl

fbs – cukier we krwi(na czczo) > 120 mg/dl ( 1-tak, 0-nie)

restecg - spoczynkowe wyniki elektrokardiograficzne( wartości 0-3)

- Wartość 0: normalna

- Wartość 1: nieprawidłowość fali ST-T (odwrócenie załamka T i / lub uniesienie odcinka ST> 0,05 mV)

- Wartość 2: wykazanie prawdopodobnego lub określonego przerostu lewej komory według kryteriów Estesa.

thalach – osiągnięte maksymalne tętno

exang – angina wywołana wysiłkiem (1-tak , 0-nie)

oldpeak – spadek ST wywołany pod czas ćwiczeń względem stanu spoczynku

slope – nachylenie spadku ST pod czas ćwiczeń

-- Value 1: upsloping   
-- Value 2: flat   
-- Value 3: downsloping

ca - liczba głównych naczyń ( wartości 0-3) zabarwionych za pomocą fluoroskopii

thal – 3 – normalny; 6 – wyleczony , 7- wada odwracalna

target – 0-zdrowy, (1-4) - chory

Jako klasa wybrana kolumna target została ona przekształcona tak aby przechowywała tylko wartości 0 i 1 gdzie 1 będzie znaczyło chory .

3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Klasyfikiator | Macierz błędu | Dokładność | TPR | FPR |
| Ctree | |  |  |  |  | | --- | --- | --- | --- | | Predicted | Real | | | |  | 0 | 1 | | 0 | 46 | 21 | | 1 | 10 | 35 | | 0.7232143 | 0.6865672 | 0.2222222 |
| kNN | |  |  |  |  | | --- | --- | --- | --- | | Predicted | Real | | | |  | 0 | 1 | | 0 | 40 | 15 | | 1 | 12 | 40 | | 0.7476636 | 0.7272727 | 0.2307692 |
| naiveByes | |  |  |  |  | | --- | --- | --- | --- | | Predicted | Real | | | |  | 0 | 1 | | 0 | 37 | 12 | | 1 | 15 | 43 | | 0.7476636 | 0.755102 | 0.2586207 |
| SVM | |  |  |  |  | | --- | --- | --- | --- | | Predicted | Real | | | |  | 0 | 1 | | 0 | 40 | 12 | | 1 | 12 | 43 | | 0.7757009 | 0.7692308 | 0.2181818 |

TP – oznacza liczbę osób, którym zgadł że są zdrowe

FP – oznacza liczbę pomylonych osób, które oznaczył jako zdrowe a były chore

FN – oznacza liczbę pomylonych osób, które oznaczył jako chore a były zdrowe

TN – oznacza liczbę osób, którym zgadł że są chore

TPR przedstawia odsetek poprawnie zdiagnozowanych(zdrowych) a TNR poprawnie zdiagnozowanych(chorych). Natomiast wartości FPR i FNR przedstawiają odsetki źle zdiagnozowanych.

Błąd pierwszego stopnia u mnie to liczba osób źle sklasyfikowanych jako chorzy a są zdrowi.

Błąd drugiego stopnia u mnie to liczba osób chorych sklasyfikowanych jako zdrowe.

Im więcej błędów pierwszego stopnia to więcej osób mogło by być niepotrzebnie leczonych.

TNR i FPR przedstawiają skuteczność testu. A FNR i FPR nieskuteczność.

W mojej bazie i sposobie modelowania gorszy był by błąd drugie stopnia, ponieważ osby chore nie były by wgl leczone.