Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Sztuczna Inteligencja

Sprawozdanie z laboratorium

Autor

Aleksandra Walczybok

nr albumu: **272454**

kierunek: **Inżynieria Systemów**

23 listopada 2024

1 Wstęp – sformułowanie problemu

Celem przeprowadzonych laboratoriów było zaimplementowanie sieci Bayesa w celu przewidywania ryzyka udaru. Wczesne wykrycie potencjalnego ryzyka udaru ma kluczowe znaczenie dla wdrożenia działań profilaktycznych i terapeutycznych, które mogą zapobiec poważnym konsekwencjom zdrowotnym lub uratować życie pacjenta.

Sieci Bayesa, będące modelem probabilistycznym opartym na teorii prawdopodobieństwa i twierdzeniu Bayesa, stanowią efektywne narzędzie do analizy danych. W kontekście problematyki medycznej, takiej jak przewidywanie ryzyka udaru, modele te pozwalają na uwzględnienie różnych zmiennych wejściowych, takich jak wiek, płeć, poziom ciśnienia krwi czy maksymalne tętno.

W ramach zadania konieczne było zaprojektowanie struktury sieci, określenie zależności między zmiennymi oraz nauczenie modelu na podstawie dostarczonych danych. Kluczowym elementem było również zrozumienie i interpretacja wyników uzyskanych z modelu, co umożliwia ocenę jego skuteczności oraz potencjalne zastosowanie w praktyce. W sprawozdaniu opisano przebieg implementacji, wykorzystane metody oraz wnioski wynikające z uzyskanych rezultatów.

2 Opis danych

Dane potrzebne do modelu pozyskałam ze strony `https://archive.ics.uci.edu/`. Dokładniej wykorzystany zbiór danych ma nazwę "Heart Disease"

Ten zbiór danych pochodzi z 1988 roku i składa się z czterech medycznych baz danych: Cleveland, Węgry, Szwajcaria i Long Beach V. Zawiera 76 atrybutów, w tym atrybut przewidywany, ale wszystkie opublikowane eksperymenty odnoszą się do wykorzystania podzbioru 14 z nich.

Są to:

1. wiek
2. płeć
3. cp - ból klatki piersiowej
   - 1 - typowa dławica piersiowa
   - 2 - dławica atypowa
   - 3 - ból inny niż dławicowy
   - 4 - bezobjawowy
4. trestbps - spoczynkowe ciśnienie krwi
5. chol - wartość cholesterolu
6. fbs - poziom cukru we krwi
7. restecg - spoczynkowe wyniki elektrokardiograficzne
   - 0 - normalna
   - 1 - występowanie nieprawidłowości załamka ST-T
   - 2 - ykazujący prawdopodobny lub określony przerost lewej komory według kryteriów Estesa
8. thalach - maksymalne tętno
9. exang - dławica wywołana wysiłkiem fizycznym (1 = tak; 0 = nie)

10. oldpeak - obniżenie odcinka ST wywołane wysiłkiem fizycznym w stosunku do odpoczynku
11. slope - nachylenie szczytowego odcinka ST podczas wysiłku
    - 0 - wznosząca się
    - 1 - płaska
    - 2 - opadanie
12. ca - liczba głównych naczyń (0-3) zabarwionych fluorosopią
13. thal
14. target (1 = ryzyko udaru; 0 = brak ryzyka)

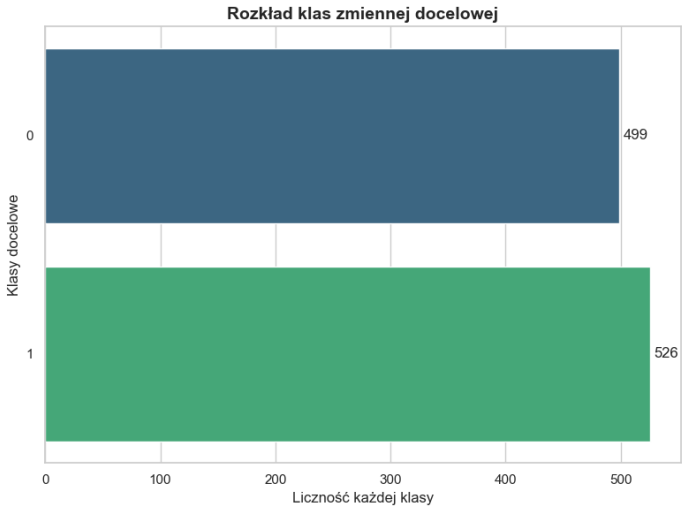
2.1 Charakterystyka danych

Tabela 1: Statystyki opisowe dla poszczególnych zmiennych

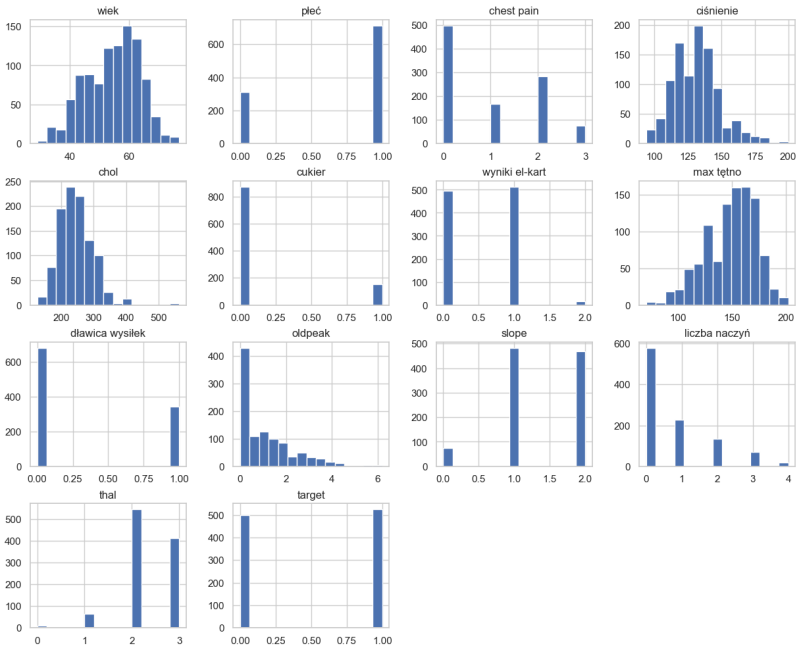
Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Age	1025.0	54.434146	9.072290	29.0	48.0	56.0	61.0	77.0
Sex	1025.0	0.695610	0.460373	0.0	0.0	1.0	1.0	1.0
Cp	1025.0	0.942439	1.029641	0.0	0.0	1.0	2.0	3.0
Trestbps	1025.0	131.611707	17.516718	94.0	120.0	130.0	140.0	200.0
Chol	1025.0	246.000000	51.592510	126.0	211.0	240.0	275.0	564.0
Fbs	1025.0	0.149268	0.356527	0.0	0.0	0.0	0.0	1.0
Restecg	1025.0	0.529756	0.527878	0.0	0.0	1.0	1.0	2.0
Thalach	1025.0	149.114146	23.005724	71.0	132.0	152.0	166.0	202.0
Exang	1025.0	0.336585	0.472772	0.0	0.0	0.0	1.0	1.0
Oldpeak	1025.0	1.071512	1.175053	0.0	0.0	0.8	1.8	6.2
Slope	1025.0	1.385366	0.617755	0.0	1.0	1.0	2.0	2.0
Ca	1025.0	0.754146	1.030798	0.0	0.0	0.0	1.0	4.0
Thal	1025.0	2.323902	0.620660	0.0	2.0	2.0	3.0	3.0
Target	1025.0	0.513171	0.500070	0.0	0.0	1.0	1.0	1.0

Tabela 2: Korelacja cech z zmienną docelową w porządku malejącym

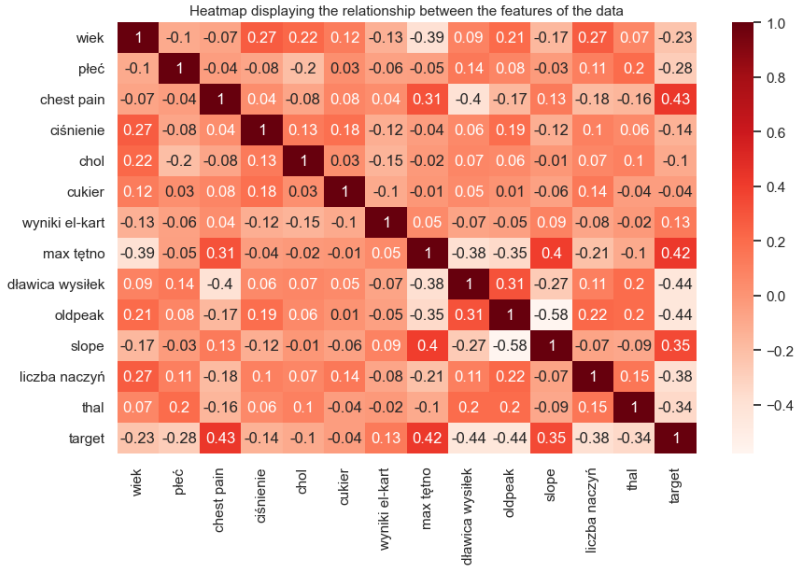
Feature	Correlation
Target	1.000000
Oldpeak	0.438441
Dławica wysiłkowa	0.438029
Ból klatki	0.434854
Max tętno	0.422895
Liczba naczyń	0.382085
Slope	0.345512
Thal	0.337838
Płeć	0.279501
Wiek	0.229324



Rysunek 1: Rozkład klasy docelowej



Rysunek 2: Histogramy zmiennych



Rysunek 3: Korelacje między zmiennymi

3 Opis rozwiązania

3.1 Wybór zmiennych

Do implementacji sieci wykorzystano 8 zmiennych. Wybór dokonano na podstawie analizy korelacji między zmiennymi. W szczególności początkowo wybrano te zmienne, które mają największą korelację z targetem. Są to: dławica, ból klatki, max tętno, liczba naczyń. Ze względu na to, że oldpeak oraz slope mają sporą korelację z max tętnem oraz dławicą postanowiono zrezygnować z tych wartości, ponieważ są one dość mocno medyczne i ich wartości mogą być ciężko dostępne. Ostatnie trzy zmienne jakie wybrano do systemu to płeć i wiek oraz ciśnienie krwi. Ostatecznie w sieci wykorzystane są zmienne:

- wiek
- płeć
- ciśnienie
- max tętno
- liczba naczyń
- dławica wysiłkowa
- ból klatki piersiowej

3.2 Przekształcenie zmiennych

Na podstawie wartości zmiennej `wiek` dokonano przypisania każdej osoby do jednej z czterech grup wiekowych. Kategoryzacja przebiegała według następujących reguł:

- **Dziecko**: jeśli $wiek < 18$,
- **Młody dorosły**: jeśli $18 \leqslant wiek < 40$,
- **Dorosły**: jeśli $40 \leqslant wiek < 60$,
- **Emeryt**: jeśli $wiek \geqslant 60$.

Jeśli chodzi o zmienną `ciśnienia` również przydzielono wartości do 4 grup, sygnalizujących czy jest to optymalna wartości nie. Kolejne stopnie mówią o stopniach zaawansowania.

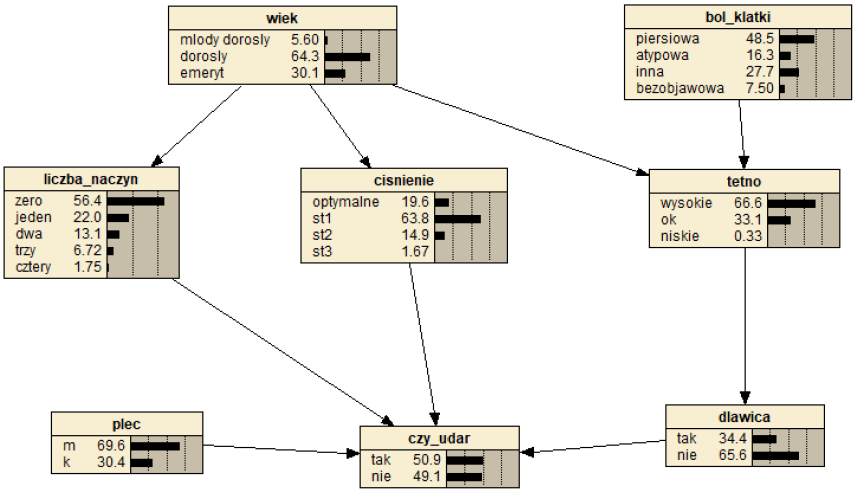
- **Optymalne**: jeśli $cisnienie < 120$,
- **Stopień 1**: jeśli $120 \leqslant cisnienie < 150$,
- **Stopień 2**: jeśli $150 \leqslant cisnienie < 180$,
- **Stopień 3**: jeśli $cisnienie \geqslant 180$.

Na tej samej zasadzie zastosowano również kategoryzację dla `max tętno`. Tutaj wykorzystano również specjalny wzór do wyliczenia odpowiedniej wartości dla każdej osoby - ponieważ max tętno jest zależne od wieku.

- **Za niskie**: jeśli $max\ tetno < 0.5 \cdot (220 - wiek)$,
- **Optymalne**: jeśli $0.5 \cdot (220 - wiek) \leqslant max\ tetno < 0.85 \cdot (220 - wiek)$,
- **Za wysokie**: jeśli $max\ tetno \geqslant 0.85 \cdot (220 - wiek)$.

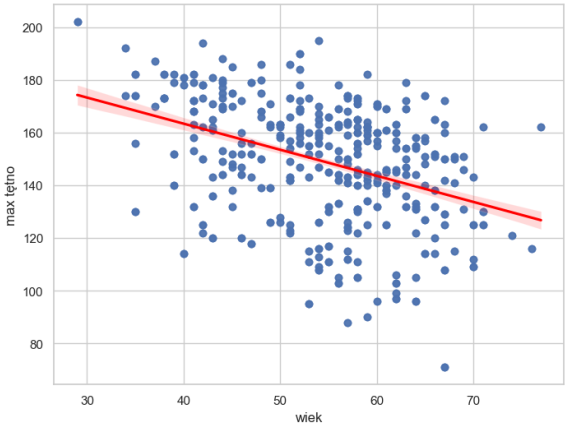
3.3 Architektura sieci

Zaimplemntowana architektura sieci wygląda w następujący sposób:



Rysunek 4: Sieć Bayesa

Zależności między zmiennymi uzyskano obserwując i analizując macierz korelacji 3 oraz wykresy takie jak poniższy.



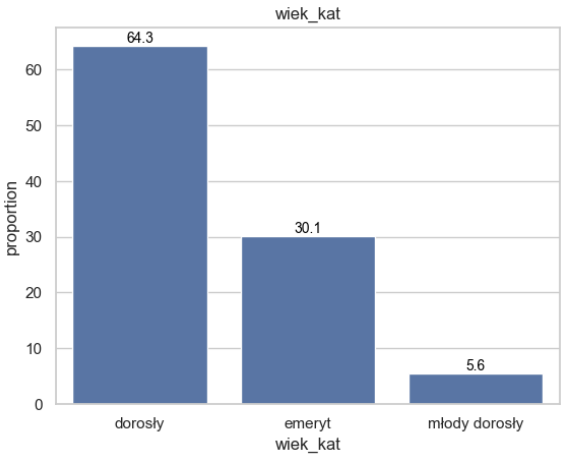
Rysunek 5: Przykładowy wykres reprezentujący zależność między zmiennymi

3.4 Obliczanie prawdopodobieństw

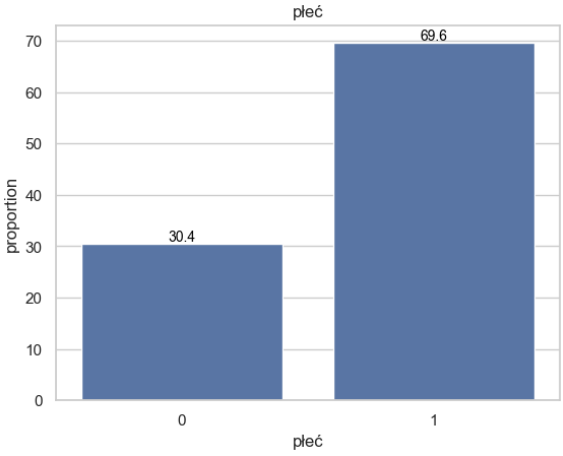
Prawdopodobieństwa jak i wcześniejsze zadania wykonano w języku Python.

Wyliczanie rozpoczęto od węzłów niezależnych od innych, czyli : płeć, wiek, ból klatki piersiowej.

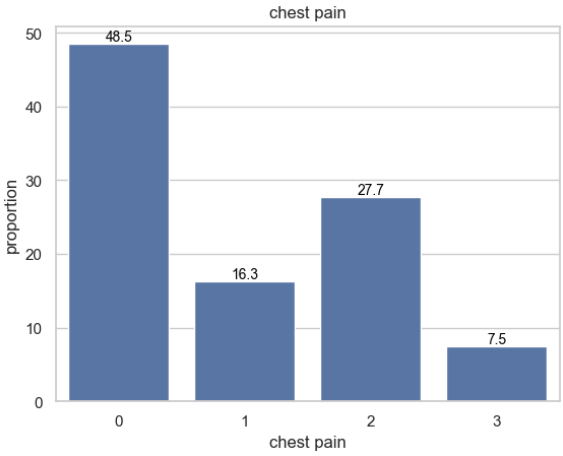
Prawdopodbieństwa do węzłów wpisywano w procentach na podstawie uzyskanych w skrypcie wykresów:



Rysunek 6: Wykres procentowy dla wiek

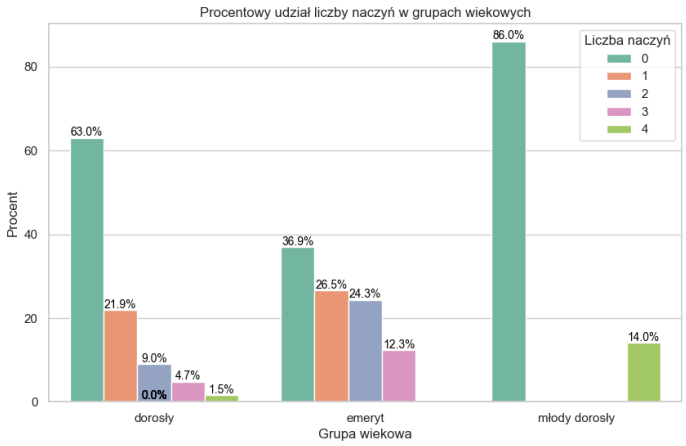


Rysunek 7: Wykres procentowy dla płeć



Rysunek 8: Wykres procentowy dla ból klatki

Prawdopodbieństwa warunkowe natomiast wyliczano w podobny sposób, grupując dane nie tylko po jednej zmiennej, ale wielu, co na wykresie przykładowo prezentowało się w następujący sposób:



Rysunek 9: Liczba naczyń procentowo w zależności od wieku

W przypadku takich zmiennych jak np. max tętno, gdzie zmienna ta zależy od 2 innych zmiennych ciężko zaprezentować to na wykresie, a więc tworzono takie tabele:

Tabela 3: Tabela kategorii wiekowych, bólu w klatce piersiowej, maksymalnego tętna i procentów

Wiek_kat	**Chest Pain**	**Max Tętno**	**Procent**
Dorosły	0	Optymalne	50.66
Emeryt	0	Optymalne	34.83
Młody dorosły	0	Optymalne	73.33
Dorosły	0	Za niskie	0.00
Emeryt	0	Za niskie	2.25
Młody dorosły	0	Za niskie	0.00
Dorosły	0	Za wysokie	49.34
Emeryt	0	Za wysokie	62.92
Młody dorosły	0	Za wysokie	26.67
Dorosły	1	Optymalne	5.22
Emeryt	1	Optymalne	45.45
Młody dorosły	1	Optymalne	0.00
Dorosły	1	Za niskie	0.00
Emeryt	1	Za niskie	0.00
Młody dorosły	1	Za niskie	0.00
Dorosły	1	Za wysokie	94.78
Emeryt	1	Za wysokie	54.55
Młody dorosły	1	Za wysokie	100.00
Dorosły	2	Optymalne	20.88
Emeryt	2	Optymalne	25.64
Młody dorosły	2	Optymalne	16.67
Dorosły	2	Za niskie	0.00
Emeryt	2	Za niskie	0.00
Młody dorosły	2	Za niskie	0.00
Dorosły	2	Za wysokie	79.12
Emeryt	2	Za wysokie	74.36
Młody dorosły	2	Za wysokie	83.33
Dorosły	3	Optymalne	23.08
Emeryt	3	Optymalne	9.68
Młody dorosły	3	Optymalne	0.00
Dorosły	3	Za niskie	0.00
Emeryt	3	Za niskie	0.00
Młody dorosły	3	Za niskie	0.00
Dorosły	3	Za wysokie	76.92
Emeryt	3	Za wysokie	90.32
Młody dorosły	3	Za wysokie	100.00

Dokładne kody prezentujące jak zostały wyliczone te prawdopodobieństwa znajdują się w załączniku.

Dla przykładu powyższa tabela powstała w taki sposób wykorzystując bibliotekę pandas:

```
# Grupowanie danych wedlug kategorii wiekowej, bolu w klatce piersiowej oraz maksymalnego tetna
grouped = data.groupby(['wiek_kat', 'chest pain', 'max t Źtno']).size().unstack(fill_value=0)

# Obliczanie procentow dla kazdej kombinacji kategorii
percentages = round((grouped.T / grouped.sum(axis=1)).T * 100, 2)

# Przywracanie indexu i przeksztalcanie tabeli do formatu dĆugiego (long format)
percentages = percentages.reset_index().melt(id_vars=['wiek_kat', 'chest pain'], var_name='max tetno', value_name='procent')
```

Listing 1: Kod w Pythonie do obliczenia procentów na podstawie grupowania danych

4 Wnioski i zastosowania sieci

Sieć Bayesa w tym przypadku została zaimplementowana w celu oceny ryzyka udaru, wykorzystując różne czynniki ryzyka, które mają istotne znaczenie w diagnozowaniu tego schorzenia. W szczególności sieć ta analizuje takie cechy, jak wiek, płeć, ciśnienie krwi, maksymalne tętno, obecność bólu w klatce piersiowej, dławicę piersiową, liczbę naczyń krwionośnych oraz inne parametry medyczne, które mogą wskazywać na potencjalne zagrożenie udarem mózgu. Dzięki zastosowaniu tej sieci możliwe jest oszacowanie prawdopodobieństwa wystąpienia udaru u pacjenta na podstawie indywidualnych wyników tych czynników.

Model pozwala również na ocenę, czy poszczególne parametry zdrowotne są w normie, czy mogą wskazywać na potencjalne ryzyko. Na przykład, analiza tętna maksymalnego w zależności od wieku pacjenta może pomóc w określeniu, czy tętno jest odpowiednie dla danej grupy wiekowej, co jest istotne w ocenie ogólnej kondycji zdrowotnej i ryzyka chorób serca. Podobnie, na podstawie wartości ciśnienia krwi, sieć może wskazać, czy poziom ciśnienia jest optymalny, co jest kluczowe w zapobieganiu udarom i innym schorzeniom układu krążenia.

A Dodatek

Kody źródłowe umieszczone zostały w repozytorium github:
`https://github.com/aleksandra0014/BAYES-NETWORKS`.