*Paweł Brodziak*

*Marceli Ptak*

*Magdalena Leśniak*

**PROJEKT SZTUCZNE SIECI NEURONOWE**

1.      W projekcie dotyczącym sztucznych sieci neuronowych podjęliśmy się problemu rozpoznawania jednostki chorobowej, jaką jest choroba wieńcowa (niewydolność) lub atak serca na podstawie danych czynników obejmujących różne parametry stanu zdrowia, jak i stylu życia. Rozwiązanie, jakie zostało stworzone pozwala na zweryfikowanie***jak bardzo prawdopodobne jest wystąpienie choroby wieńcowej/ataku serca.***

Projekt wykonaliśmy w języku Python.   
Dane pozyskaliśmy z ogólnodostępnego serwisu dysponującego różnego rodzaju zbiorami danych. Pierwotnie zestaw ten obejmował więcej zmiennych, natomiast zdecydowaliśmy się wyeliminować te mniej znaczące i ograniczyć nasze dane do ilości 14 kolumn - według nas zawierają one najbardziej istotne czynniki.

Większość zmiennych przyjmuje warunek zero-jedynkowy, poniżej krótkie objaśnienie, jak należy je interpretować:

* HighBP – czy kiedykolwiek pracownik służby zdrowia zadeklarował u pacjenta nadciśnienie
* HighChol – czy kiedykolwiek pracownik służby zdrowia zadeklarował u pacjenta wysoki poziom cholesterolu
* CholCec – kontrola cholesterolu pacjenta w ciągu ostatnich pięciu lat
* BMI – wskaźnik masy ciała pacjenta
* Smoker – czy w całym życiu pacjent wypalił co najmniej 100 papierosów
* Stroke – czy kiedykolwiek pacjent miał udar
* Diabetes – czy kiedykolwiek zadeklarowano pacjentowi, że ma cukrzycę
* PhysActivity – czy w ciągu ostatnich 30 dni pacjent wykonywał aktywność fizyczną inną niż praca
* Fruits – czy pacjent spożywa owoce 1 lub więcej razy dziennie
* Veggies – czy pacjent spożywa warzywa 1 lub więcej razy dziennie
* HvyAlcoholConsump – czy pacjent pije dużo alkoholu (więcej niż 14 drinków mężczyzna/7 kobieta)
* DiffWalk – czy pacjent ma poważne trudności z wchodzeniem po schodach
* Sex – płeć, 1 mężczyzna, 0 kobieta
* Age – według 14-poziomowej kategorii wiekowej (18-24, 25-29 itd. (zakres 4 lata)).

Dane podzieliliśmy na zbiór uczący się (80%) i testowy (20%).

Sieć w naszym projekcie należy do typu jednokierunkowego - MLP - perceptronu wielowarstwowego. Do wykonania sieci wykorzystaliśmy łącznie cztery warstwy – wejściową, dwie ukryte oraz wyjściową. Warstwa wejściowa obejmowała 14 neuronów, wyjściowa 1, natomiast warstwy ukryte po 6 neuronów – w naszej opinii rozwiązanie takie było optymalne. Obliczenie wartości wyjścia neuronów zostało wyznaczone na podstawie funkcji sigmoidalnej, użyte również były metody propagacji w przód, wstecznej propagacji, algorytm spadku wzdłuż gradientu.

2.      Wiedzę na temat sztucznych sieci neuronowych pogłębialiśmy z szeroko dostępnych opracowań powyższego zagadnienia dostępnych w zasobach Internetu. Między innymi bazowaliśmy na poniższych pozycjach:

* Heaton Research – witryna zawierająca artykuły Jeffa Heatona, autora *Introduction to Neural Networks for Java*, w której prezentował między innymi strukturę/architekturę sieci,
* Leksykon Sieci Neuronowych - Ryszard Tadeusiewicz, Maciej Szaleniec – pozycja zawierająca ogólne przedstawienie tematu sieci neuronowych
* A DATA MINING APPROACH FOR PREDICTION OF HEART DISEASE USING NEURAL NETWORKS Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte – artykuł badawczy, w którym przedstawiono system przewidywania chorób serca za pomocą sztucznych sieci neuronowych z użyciem algorytmu wstecznej propagacji błędu.
* System do wspomagania stawiania diagnoz chorób serca. B. Hyży, M. Pieróg, W. Baran, J. Jaśkowiec, M. Niesyto - projekt dotyczy informatycznego wspomagania lekarzy przy stawianiu diagnoz, zbudowany został w oparciu o sieci neuronowe, a obejmowany przez niego zakres to choroby serca.

3.      Analiza czynników:

3.1.  Liczba warstw - zwykle sieci składają się z minimum trzech warstw - z czego problematycznym elementem jest kwestia ilości warstw ukrytych. W przypadku braku warstwy ukrytej, sieć dzieli przestrzeń sygnałów wejściowych na tylko dwie części. W przypadku pojedynczej warstwy ukrytej, obszar decyzyjny ulega zmianie i przyjmuje kształt wypukłego wielokąta (simpleks). Dwie warstwy ukryte pozwalają wygenerować dokładniejsze kształty obszaru decyzyjnego, wliczając w to również powierzchnie niewypukłe, co z reguły jest wystarczające. W naszym projekcie zastosowaliśmy dwie warstwy ukryte - pomimo iż problem nie był bardzo zaawansowany, jak i fakt, że dysponujemy prostymi, nieskomplikowanymi danymi i warstwa wyjściowa nie jest bardzo złożona - takie rozwiązanie pozwoli na zapewnienie dobrej jakości działania sieci.

3.2. Liczba neuronów - jeśli chodzi o warstwę ukrytą - duża ilość neuronów sprawia więcej kłopotów podczas uczenia, natomiast niedostateczna ilość jest zbyt prymitywna, aby poradzić sobie z trudniejszymi zadaniami. Stosowane metody na określenie ilości neuronów to: *pierwiastek z iloczynu liczby neuronów wejścia i wyjścia* lub *suma połowy ilości neuronów wejścia i neuronów wyjścia* lub *⅔ ilości neuronów wejścia*. Generalnie sieci o większej ilości neuronów w warstwie ukrytej radzą sobie lepiej, natomiast zbyt duża ilość takich neuronów powoduje, że sieć jest “nadmiernie inteligentna” (problemy z uczeniem się). Zgodnie z powyższymi zasadami przyjęliśmy średnio po 6 neuronów w każdej warstwie ukrytej względem 14 w warstwie wejściowej.

3.3. Wielkość i sposób doboru grupy uczącej się i testowej - dane podzieliliśmy w proporcji 80/20, z czego większość stanowi część ucząca się. Głównym elementem związanym z podziałem na grupy jest konieczność wystąpienia w zbiorze uczącym się przypadków reprezentatywnych - nie mogą być one unikatowe, jak również nie mogą być zbyt standardowe, ponieważ ocena jakości sieci nie będzie skuteczna. Ze względu na prostotę posiadanych danych, jak i ich dużą liczebność, uznaliśmy, że podział 80/20 będzie optymalnym rozwiązaniem.