Tomasz Szewczuk, 226163
Adam Węglowski, 226175

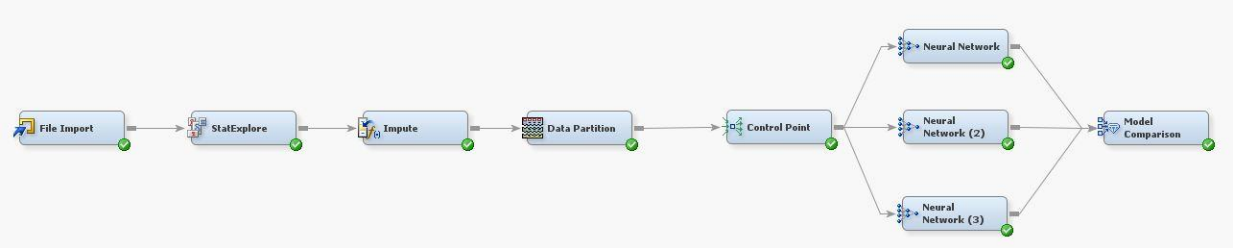
Zaawansowane metody analizy danych medycznych

Projekt – część II

Prowadzący:

Dr hab. inż. Robert Burduk

Wybrany zbiór danych, w którym zebrane zostały wszystkie przeprowadzone badania pochodzi z serwisu Kaggle. Zbiór danych dotyczący chorób serca pochodzi z 1988 roku. Złożony jest z czterech zestawów danych przebadanych pacjentów: Cleveland i Long Beach V – miasta w Stanach Zjednoczonych, Węgry oraz Szwajcaria. Plik heart.csv został załadowany do programu poprzez blok *File Import*.



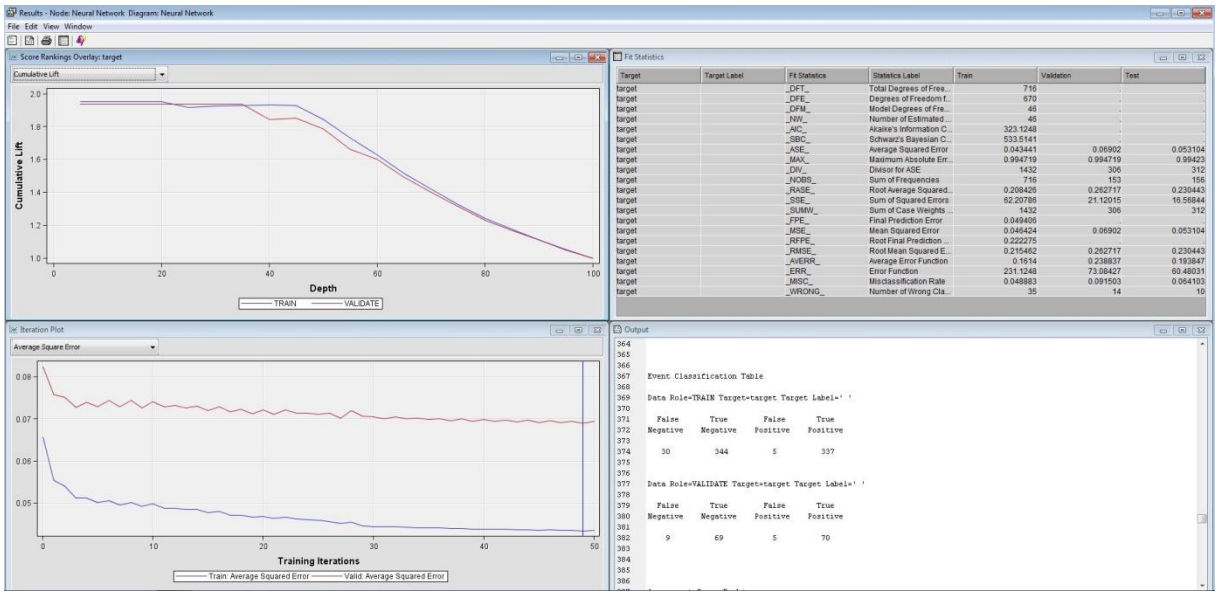
Rysunek 1. Diagram stworzony w programie Enterprise Miner

Kolejny blok o nazwie *StatExplore* odpowiadał za przekazanie informacji o zbiorze danych. Natomiast *Impute* to bloczek mający za zadanie ewentualne uzupełnienie brakujących danych. Dane posiadają 14 atrybutów, w tym etykietę wyniku przewidywanego „target". Kolumna „target" zawiera informacje o rozpoznaniu przypadku. Wartość 1 oznacza chorobę, natomiast wartość 0, to brak choroby. Liczba wykonanych badań dostępnych w zbiorze to 1025, z czego 526 to przypadki chorób serca, a 499 to próbki, które okazały się badaniami osób zdrowych. Z powyższych liczb, odnośnie ilości przypadków można wnioskować, że wybrany zbiór danych jest dobrze zbalansowany. Oznacza to że, jedna klasa wynikowa znacząco nie góruje nad drugą klasą, jeśli chodzi o ilość badanych próbek. Wybrany zbiór danych nie posiada wartości nieznanych czy też pustych.

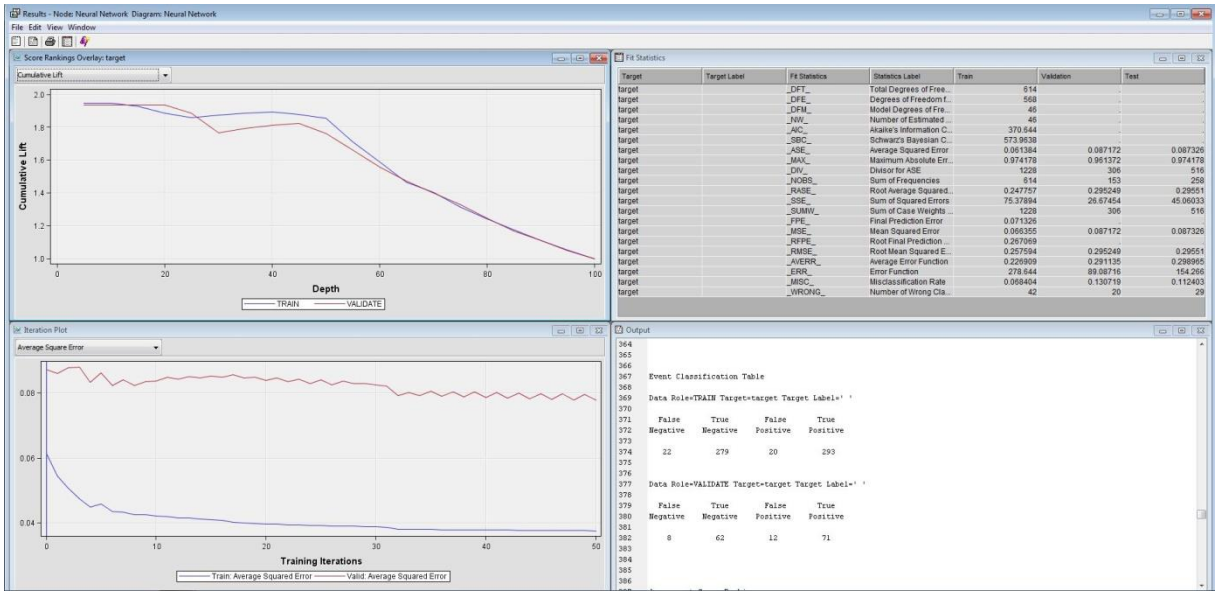
Ważnym elementem jest *Data Partition*, za jego pomocą wykonano podział danych wejściowych na zbiory: uczący, testowy i walidacyjny.

Wybranym klasyfikatorem, dla którego przeprowadzone zostały badania eksperymentalne, była sieć neuronowa.

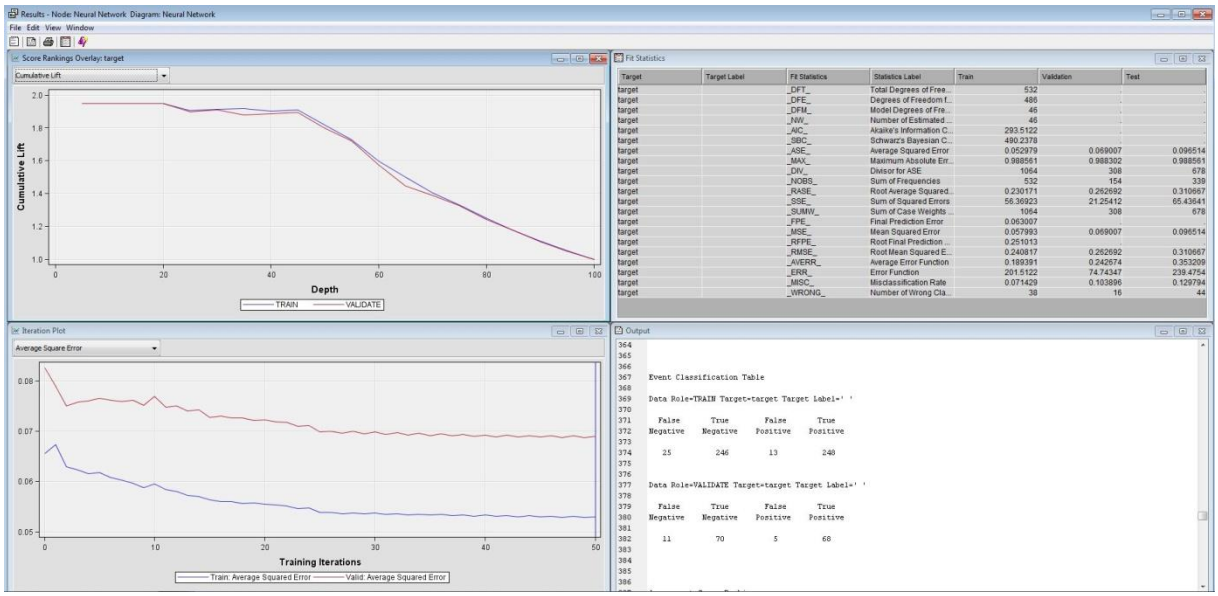
Testy zostały wykonane dla odpowiednio trzech, czterech, pięciu oraz sześciu jednostek ukrytych w warstwach sieci neuronowej, dla wyznaczonych podziałów procentowych zbioru wejściowego. Zbiór walidacyjny został określony w każdym przypadku na 15% początkowego zbioru. Parametrem był procent zbioru testowego (odpowiednio 15%, 25%, 33%). Liczba epok uczenia została ustalona na wartość 50.



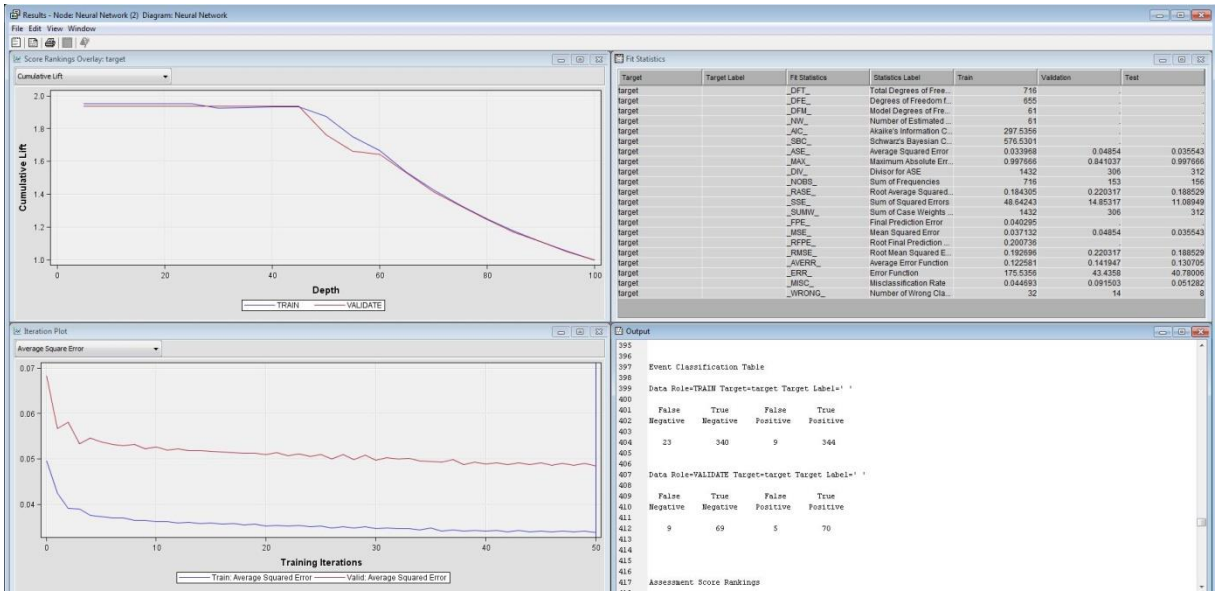
Rysunek 2. Wyniki dla jednostek ukrytych = 3 oraz podziału zbioru testowego 15%



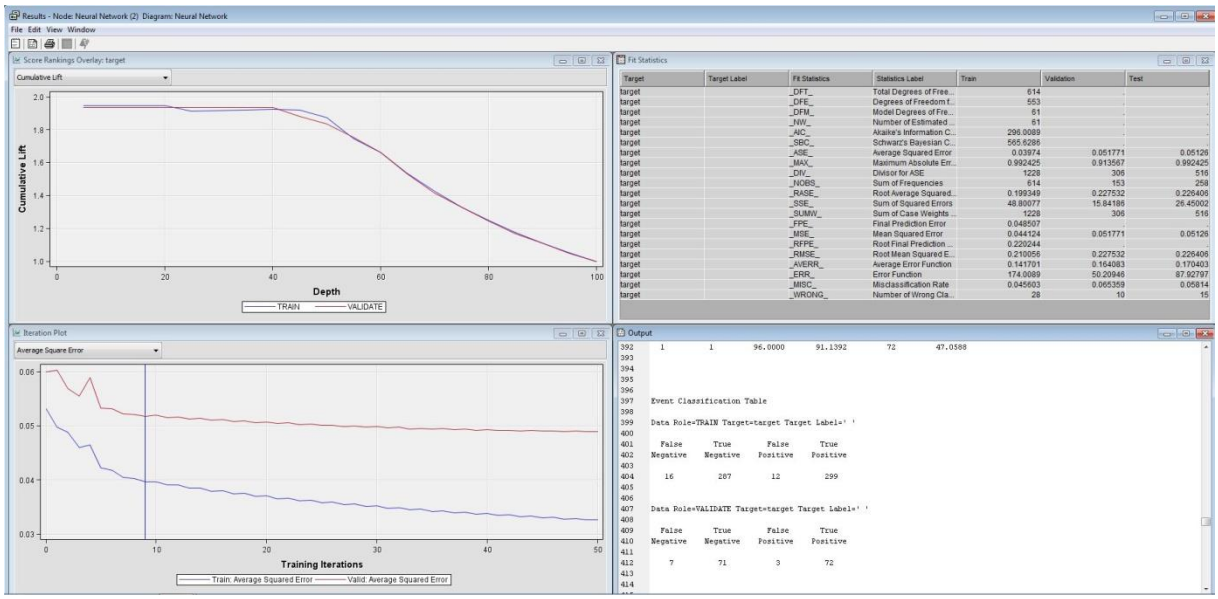
Rysunek 3. Wyniki dla jednostek ukrytych = 3 oraz podziału zbioru testowego 25%



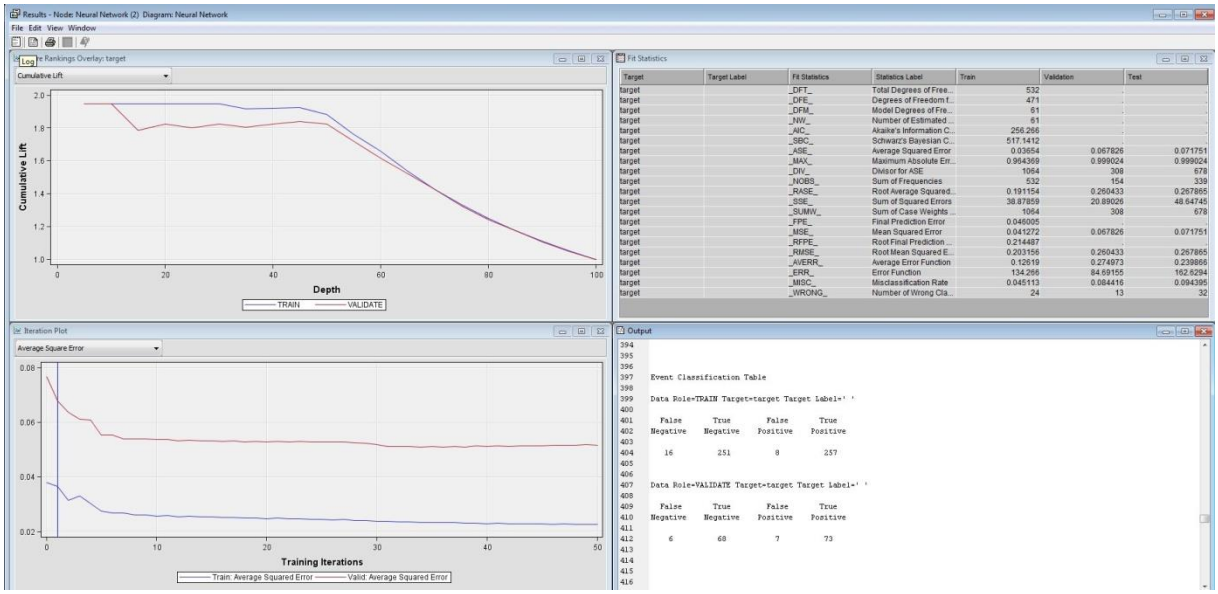
Rysunek 4. Wyniki dla jednostek ukrytych = 3 oraz podziału zbioru testowego 33%



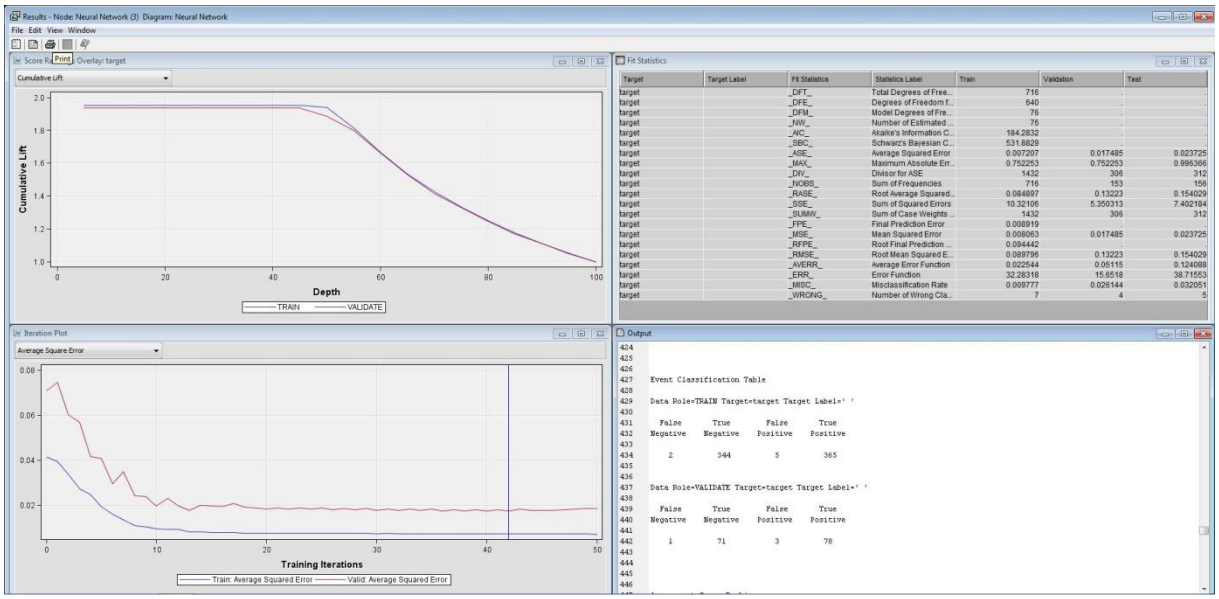
Rysunek 5. Wyniki dla jednostek ukrytych = 4 oraz podziału zbioru testowego 15%



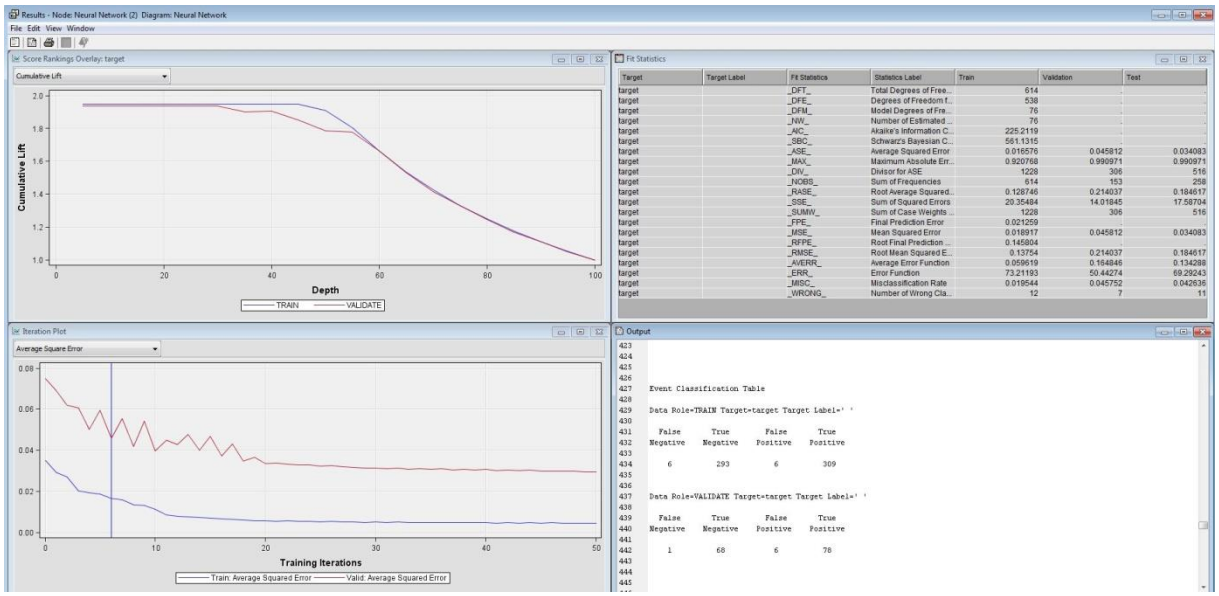
Rysunek 6. Wyniki dla jednostek ukrytych = 4 oraz podziału zbioru testowego 25%



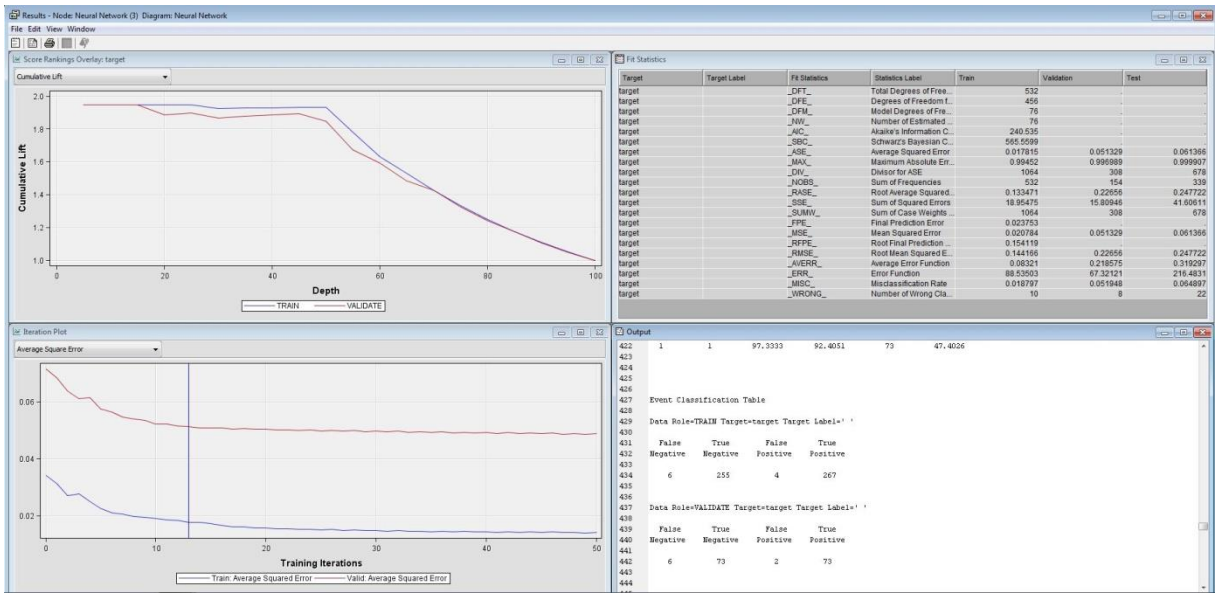
Rysunek 7. Wyniki dla jednostek ukrytych = 4 oraz podziału zbioru testowego 33%



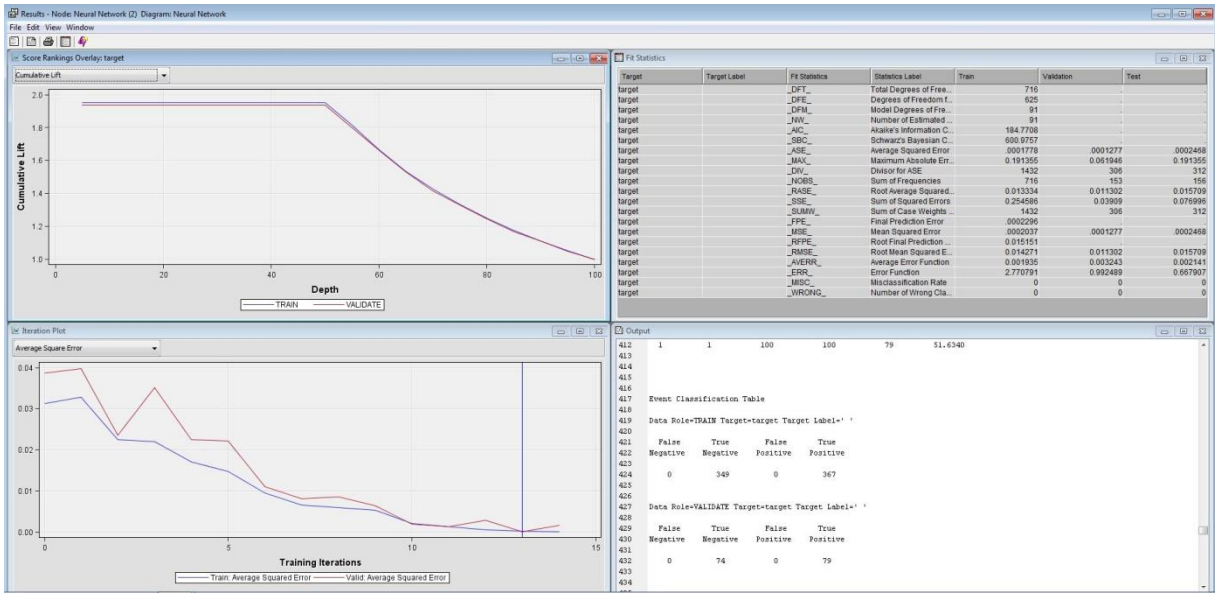
Rysunek 8. Wyniki dla jednostek ukrytych = 5 oraz podziału zbioru testowego 15%



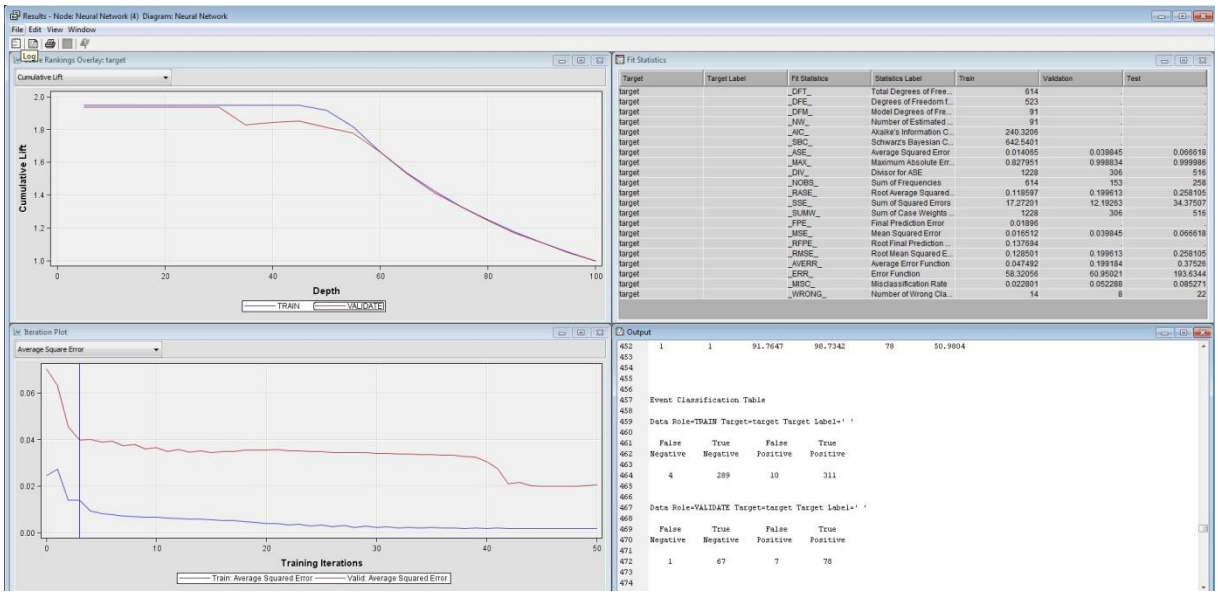
Rysunek 9. Wyniki dla jednostek ukrytych = 5 oraz podziału zbioru testowego 25%



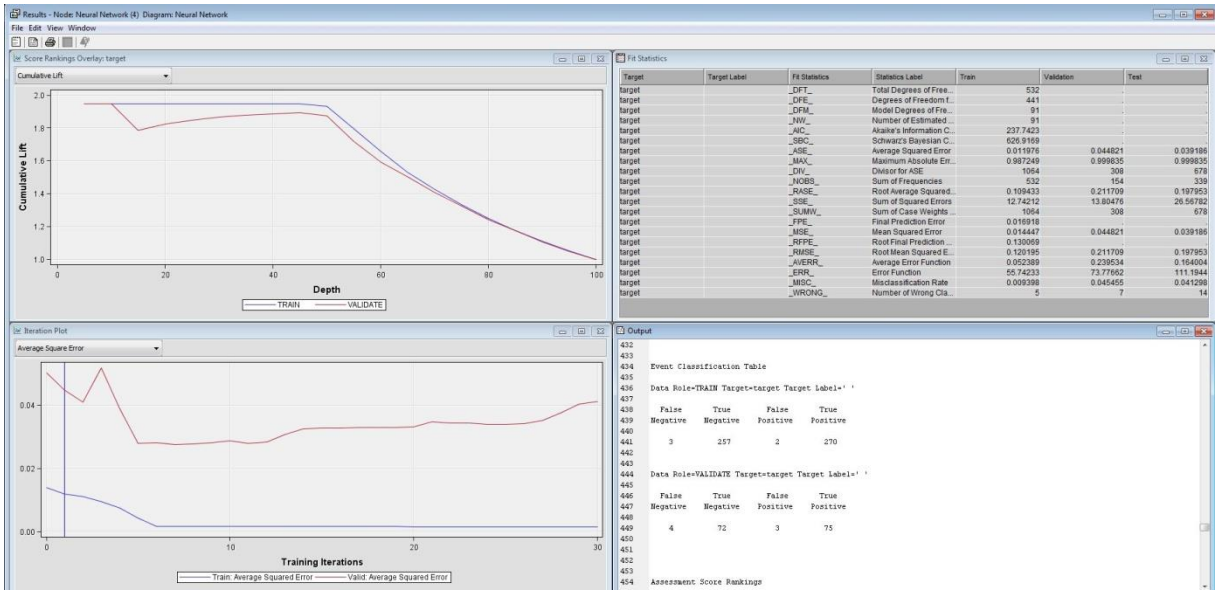
Rysunek 10. Wyniki dla jednostek ukrytych = 5 oraz podziału zbioru testowego 33%



Rysunek 11. Wyniki dla jednostek ukrytych = 6 oraz podziału zbioru testowego 15%



Rysunek 12. Wyniki dla jednostek ukrytych = 6 oraz podziału zbioru testowego 25%



Rysunek 13. Wyniki dla jednostek ukrytych = 6 oraz podziału zbioru testowego 25%

Wyniki poszczególnych badań zostały ukazane w tabeli poniżej, gdzie główną metryką porównawczą była dokładność. Dokładność jest to stosunek liczby poprawnie sklasyfikowanych próbek do liczby wszystkich próbek dostępnych w bazie.

Tabela 1. Wyniki przeprowadzonego eksperymentu w Enterprise Miner

jedn. ukryte / % zbioru	3 jednostki	4 jednostki	5 jednostek	6 jednostek
15%	95,11%	95,53%	99,02%	100%
25%	93,16%	95,44%	98,05%	97,72%
33%	92,86%	95,49%	98,12%	99,06%

Dodatkowo dla porównania przeprowadzone zostały również badania w środowisku programistycznym Python. Sieć neuronowa została zbudowana za pomocą biblioteki dla języka Python o nazwie Keras. Keras jest pakietem, dzięki któremu łatwo można definiować i trenować dowolne modele uczenia maszynowego.

Tabela 2. Wyniki przeprowadzonego eksperymentu w środowisku Python

jedn. ukryte / % zbioru	3 jednostki	4 jednostki	5 jednostek	6 jednostek
15%	96,75%	96,05%	98,70%	99,35%
25%	93,39%	95,50%	97,67%	99,22%
33%	92,33%	95,57%	95,57%	98,82%

Zgodnie z uzyskanymi wynikami można wyciągnąć wnioski, że im większa liczba jednostek ukrytych w poszczególnych warstwach sieci neuronowej, tym znacząco lepsze wyniki klasyfikacji są uzyskiwane. Analizując badania dla parametru procentowego podziału zbioru testowego zauważyć można, że wyniki dla 15% podziału w każdym przypadku są lepsze od pozostałych badanych podziałów. Taki stan rzeczy jest spowodowany przez zbyt małą ilość próbek w zbiorze uczącym dla innych podziałów.

Ponadplanowo w projekcie porównano dwie platformy badawczych tzn. środowisko Enterprise Miner oraz własną implementację w języku Python. Zgodnie z wynikami przedstawionymi w Tabeli 1 oraz Tabeli 2 nie można jednoznacznie stwierdzić, które ze środowisk poradziło sobie znacząco lepiej z problemem klasyfikacji chorób serca. Obie metody uzyskały satysfakcjonujące wyniki klasyfikacji (ponad 90% dokładności).