Znalezienie odpowiednich danych jest zadaniem trudnym i łatwo można popełnić błąd. W Internecie można znaleźć wiele zbiorów jednak wiele z nich nie spełnia wymagań. W pierszym kroku odrzuciłem wszystkie bazy z ilością rekordów mniejszą niż 1000 i takie które zawierały powyżej 20 atrybutów. Na dodatek należy uważać gdyż niektóre bazy zawierają dane wygenerowane według wzorów. Ostatecznie zdecydował na znaną mi z poprzedniego zadania baza abalone (http://archive.ics.uci.edu/ml/datasets/Abalone). Zawiera ona 4177 rekordy i następujące atrybuty:

Length / continuous / mm / Longest shell measurement

Diameter / continuous / mm / perpendicular to length

Height / continuous / mm / with meat in shell

Whole weight / continuous / grams / whole abalone

Shucked weight / continuous / grams / weight of meat

Viscera weight / continuous / grams / gut weight (after bleeding)

Shell weight / continuous / grams / after being dried

Rings / integer / -- / +1.5 gives the age in years

Oraz płeć która może przyjmować wartość M, F, I i właśnie ten atrybut będzie przedmiotem naszego zainteresowania. Będziemy starali się nauczyć komputer aby prawidłowo rozpoznał płeć na podstawie pozostałych atrybutów. Jak widać wszystkie parametry są ciągłe i stosunkowo równomiernie rozłożone. W bazie występują porównywalne ilości rekordów z klasą M, F oraz I.

Pierwszym krokiem było zastosowanie naiwnego klasyfikatora Bayes-a dla całego zbioru i wszystkich atrybutów. Otrzymano dokładność 52.09%. Następnie zastosowano walidacje krzyżową z wynikiem 52.00% +/- 1.55%.

Kolejno następne podzielono zbiór danych na podzbiór uczący i testujący w celu znalezienia minimalnej ilości próbek potrzebniej do nauczenia sieci.

I tak dla różnych podziałów zbioru uczącego do testowanego otrzymano następujące wyniki:

90/10 - 51.41% +/- 2.21%

70/30 - 51.62% +/- 1.19%

50/50 - 51.90% +/- 0.86%

10/90 - 51.82% +/- 0.81%

1/99 - 51.52% +/- 1.38%

0.1/99.9 - 34.63% +/- 5.22%

Dokładne wyniki wykazały iż potrzeba 0.5% próbek czyli 20 aby otrzymać wynik powyżej 50%.

Następnie poprzez dyskretyzację i wybór parametrów udało się zwiększyć dokładność do 53.63% +/- 1.78%. W procesie dyskretyzacji użyłem narzędzia pozwalającego na automatyczną dyskretyzację atrybutów ze względu na poziom ich entropii. Wybór atrybutów padł na Diameter, Rings, Viscera weight.

Wypróbowano również inne metody z następującym skutkiem

Decision Tree - 53.27% +/- 1.48%

k-NN - 47.57% +/- 2.49%

ID3 - 53.20% +/- 1.65%

NN - 55.81% +/- 1.93%

Niestety