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 adults (<http://archive.ics.uci.edu/ml/machine-learning-databases/adult/>). Zawiera ona 32tysiące rekordów i następujące atrybuty:

- wiek

- sektor zatrudnienia

- poziom edukacji

- czas edukacji

- status małżeński

- stanowisko w pracy

- status rodziny

- rasa

- płeć

- ilość godzin pracy tygodniowo

- kraj pochodzenia

Oraz płacę która może przyjmować wartość „>50K” i „<=50K”, i właśnie ten atrybut będzie przedmiotem naszego zainteresowania. Będziemy starali się nauczyć komputer aby prawidłowo rozpoznał płacę na podstawie pozostałych atrybutów.

Pierwszym krokiem było zastosowanie naiwnego klasyfikatora Bayes-a dla całego zbioru i wszystkich atrybutów. Otrzymano dokładność 81.19%. Następnie zastosowano walidacje krzyżową(81.12% +/- 0.29%). W celu poprawy wyników zastosowano dyskredytacje oraz redukcje parametrów(doświadczenie z poprzedniego zadania wykazało ze rasa nie ma wpływu na jednak wynik udało się poprawić jedynie do poziomu 82.06% +/- 0.45%. Podobne badania dokonano dla drzewa decyzyjnego otrzymując wynik 77.42% +/- 0.16%.

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