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 bazę danych z wyborem metody antykoncepcji (http://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice). Zawiera ona 1473 rekordy i następujące atrybuty:

1. Wife's age (numerical)

2. Wife's education (categorical) 1=low, 2, 3, 4=high

3. Husband's education (categorical) 1=low, 2, 3, 4=high

4. Number of children ever born (numerical)

5. Wife's religion (binary) 0=Non-Islam, 1=Islam

6. Wife's now working? (binary) 0=Yes, 1=No

7. Husband's occupation (categorical) 1, 2, 3, 4

8. Standard-of-living index (categorical) 1=low, 2, 3, 4=high

9. Media exposure (binary) 0=Good, 1=Not good

Oraz metodę antykoncepcji która może przyjmować wartość 1=No-use, 2=Long-term, 3=Short-term i właśnie ten atrybut będzie przedmiotem naszego zainteresowania. Będziemy starali się nauczyć komputer aby prawidłowo rozpoznał metodę na podstawie pozostałych atrybutów. Atrybuty są stosunkowo równomiernie rozłożone. W bazie występują porównywalne ilości rekordów z każdej z klas.

Pierwszym krokiem było zastosowanie naiwnego klasyfikatora Bayes-a dla całego zbioru i wszystkich atrybutów. Otrzymano dokładność 47.45% Następnie zastosowano walidacje krzyżową z wynikiem 47.05% +/- 6.34%

Następnie poprzez dyskretyzację i wybór parametrów udało się zwiększyć dokładność do 55.40% +/- 3.80%. W procesie dyskretyzacji użyłem narzędzia pozwalającego na automatyczną dyskretyzację atrybutów ze względu na poziom ich entropii. Próby wykazały że usunięcie atrybutów „edukacja męża” i „wystawienie na media” poprawia wynik.

Wypróbowano również inne metody z następującym skutkiem

Decision Tree - 42.70% +/- 0.22%

k-NN - 43.45% +/- 1.22%

ID3 - 56.56% +/- 4.07%

NN - 53.03% +/- 3.21%

Nieznacznie lepsza okazała się metoda ID3 i do dalszych testów została wybrana właśnie ona.

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 - 55.29% +/- 3.84%

70/30 - 55.33% +/- 2.01%

50/50 - 54.64% +/- 1.62%

20/80 - 53.08% +/- 1.53%

10/90 - 51.13% +/- 2.02%

1/99 - 43.49% +/- 3.76%

Dokładne wyniki wykazały iż potrzeba 7% próbek czyli około 100 aby otrzymać wynik powyżej 50%.

Podsumowując uzyskany wynik jest daleki od doskonałości jednakże wciąż lepszy niż zgadywanie(33%). Poprzez odpowiednie manipulacje udało się poprawić wynik o prawie 10 punktów procentowych. Można przypuszczać iż tak niski wynik spowodowany jest faktem że wybór metody antykoncepcji może być podyktowany innymi czynnikami niż znane nam atrybuty.