Drugi etap: Przygotowanie danych + Modelowanie

W drugim etapie wiele zależy od wybranych przez Państwa celów eksploracji i specyfiki zbioru danych, dlatego bardzo proszę przede wszystkim kierować się zdrowym rozsądkiem!

Poniższy spis prezentuje zbiór czynności, które zwykle są wykonywane w projektach analizy danych. Tym samym, nie wszystkie czynności są obligatoryjne, zaś sama lista może być fakultatywnie uzupełniona.

Sprawozdanie z projektu winno w nagłówku zawierać Państwa dane osobowe (imię, nazwisko, numer indeksu, data, nazwa sprawozdania). Następnie, poniżej, należy wskazać:

- charakterystykę zbioru danych wejściowych,
- cel lub cele eksploracji danych,
- dyskusję kroków dalszego postępowania:
  - dobór działania eksploracji,
  - dobór algorytmu eksploracji,
  - dobór metody testowania wyników,
- przygotowanie danych, np. poprzez:
  - zdecydowanie, co zrobić z danymi brakującymi,
  - sprawdzenie, czy jakichś wartości (najczęściej tekstowych) nie należy ujednolicić,
  - zastanowienie się, czy nie należy wybrać jakiegoś podzbioru danych, bądź nie należy danych uzupełnić (np. bardzo mało przykładów pozytywnych/negatywnych),
  - zależnie od wybranej metody modelowania zastanowienie się, czy nie należy atrybutów zmienić na numeryczne/nominalne,
- utworzenie modelu:
  - dobranie parametrów pracy algorytmu,
  - wykonanie algorytmu,
  - przeanalizowanie samego modelu (np. zbadanie kształtu drzewa decyzyjnego, sprawdzenie centrów i wariancji klastrów itd.),
  - ocena wyników według dobranej metody,
- eksperymenty z modelem i zbiorem danych, np.:
  - wybranie innego podzbioru atrybutów,
  - przekształcenie niektórych atrybutów,
  - dobór innych parametrów pracy algorytmu,
  - sprawdzenie, jak eksperymenty wpływają na utworzony model i jego ocenę,
  - sprawdzenie algorytmów alternatywnych.
- krótkie podsumowanie:
  - krótki przegląd wykonanego procesu (jakie kroki przyniosły dobre rezultaty, co można byłoby ewentualnie zrobić lepiej),
  - stopień pokrycia celów.